AD_____

Award Number: DAMD17-97-1-7202

TITLE: Investigation of Genetic Algorithms for Computer-Aided Diagnosis

PRINCIPAL INVESTIGATOR: Matthew A. Kupinski

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: October 2000

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Final (1 Oct 97 - 30 Sep 00)

4. TITLE AND SUBTITLE
Investigation of Genetic Algorithms for Computer-Aided Diagnosis

5. FUNDING NUMBERS
DAMD17-97-1-7202

6. AUTHOR(S)
Matthew A. Kupinski

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

E-MAIL:
m-kupinski@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Computer-aided diagnosis has the potential of substantially increasing diagnostic accuracy in mammography. Using a computer to double-check a radiologist's findings is becoming more popular and more important as the public learns that the best defense against breast cancer is early detection. The University of Chicago is currently developing computerized schemes to detect cancers in digital mammograms. We use a pattern classification system known as an artificial neural network (ANN) to classify certain regions of the digital mammograms as cancerous or non-cancerous. ANNs are trained pattern classification devices that take, as inputs, features extracted from regions in the mammograms and output the classification. Currently, there are a total of 42 features extracted from the various regions in each mammogram. A subset of those 42 features must be chosen as inputs for the ANN. The goal of this research was to investigate methods of feature selection and pattern classification in order to improve upon the overall performance of CAD systems.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
98

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

___ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

___ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

___ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

___ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

___ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Matthew A. Kupinski 10/2/00

PI - Signature Date

Table of contents

4 Introduction

Breast cancer is a major cause of death among women over the age of forty [1]. Mammography is the most effective diagnostic procedure for the early detection of breast cancer [2, 3]. Mammography is not, however, perfect. Between 10-30% of women who have breast cancer and undergo mammography have negative mammograms [4–7]. Of these, radiologists have determined, retrospectively, that two-thirds of the cancers could have been detected [5, 6, 8, 9]. One possible means by which to decrease this number is to have two radiologists read the mammograms. This method has been shown to increase sensitivity by as much as 15%, [10, 11] but can be costly both financially and with respect to time. A computer-aided diagnostic scheme may act as an inexpensive second reading method. The final decision would be made by the radiologist.

This research sought to answer questions that arise when using pattern classifiers in decision making applications. Problems occur when the number of inputs to the pattern classifier become large. For this reason, genetic algorithms and other *feature selection* techniques were studied to alleviate this problem. The purpose of this research was to study and develop feature selection and pattern classification methods to improve the performance of CAD schemes. Specific emphasis was placed on using the developed methods in the computerized detection of mass lesions in mammography.

5 Body

5.1 Technical Objectives

The objectives of this project were as follows:

- Development of a genetic algorithm for the optimization of artificial neural network inputs.
- Comparison of a genetic algorithm with other selection and optimization methods including previously used selection methods.
- Analysis of features selected by genetic algorithm and comparison of those features with visual techniques employed by radiologists.
- Development of a parallel genetic algorithm to improve performance of the search and to provide an even greater performance increase to the mass detection CAD program.

5.2 Methods

5.3 Development of Classification Methods

We have performed extensive analysis of two methods of pattern classification. First, we studied the ability of a classifier called a Bayesian ANN to approximate the ideal observer (the theoretically optimal classifier) given datasets of limited size. A full analysis of this is given in reference [12] which is attached as Appendix A.

We also developed a novel classifier training strategy which directly optimizes the ROC curve of a given classifier on a given training dataset. This approach to classifier training was found to be useful for "simple" classifiers such as rule-based classifiers. A full analysis of this method is given in reference [13] which is attached as Appendix B.

We have applied various feature selection methods to select subsets of features for both of these classifiers using data extracted from a database 177 screening mammography cases. Validation results were computed and used to compare the various classifiers and methods of feature selection as is discussed later.

5.3.1 Development of Feature Selection Methods

We have studied various feature selection methods for selecting features to be used within an artificial neural network (ANN) classifier and methods for selecting features for rule-based classifiers. A detailed summary of these methods can be found in references [14] which is also attached as Appendix C.

We also endeavored to use these methods to select features for use in an actual computerized detection system. A summary of these results is shown in Appendix D which is a paper under review at Medical Physics. In this paper, it is shown that the Bayesian ANN outperforms a rule-based classifier trained using the multiobjective approach. Furthermore, both genetic algorithm feature selection and a forward selection method performed best for selecting features for Bayesian ANNs.

5.3.2 Analysis of Feature Selection

During the course of our research on pattern classification and feature selection, we studied in more detail, some of the fundamental problems associated with feature selection [15]. A paper on this subject is attached as Appendix E. It was found that the ability of a feature selection method to select the optimal subset of features decreases as the number of feature from which to select increases and as the database used to select the features decreases in size. For a simple feature selection problem, we derived the probability that the optimal subset of features will be selected.

5.3.3 Analysis of Features

A detailed analysis of the features selected by the various feature selection methods is also contained in Appendix D. The features we have found to be most useful in distinguishing between malignant mass lesions and false candidates in mammography were gradient-based features such as RGI [16] and intensity-based measures such as the contrast. The geometric features were found to be useful but not to the extent shown in previous studies [17]. This is due to a novel lesion segmentation technique that we have developed for the mass detection CAD system which always returns lesion-like results even in non-mass image areas [16].

5.3.4 Development of a Parallel Genetic Algorithm

A group at Argonne National labs has developed a parallel genetic algorithm package called the PGAPack [18]. This package was found to be suitable for the implementation of GA feature selection.

6 Key Research Accomplishments

- Development of an RGI-based lesion segmentation method which outperforms previously studied lesion segmentation methods.

  Development of a model-based probabilistic segmentation method which outperforms previously studied lesion segmentation methods.

- Analyzed the ability of Bayesian ANNs to approximate the ideal observer given simulated datasets. These results aided in the design of Bayesian ANN classifiers for CAD systems.

- Development of a multiobjective classifier training methodology which directly optimized the ROC curve of a classifier. This approach has been found to be very useful for rule-based classifiers.

- Analysis of the fundamentals of the feature selection problem which helps one better understand the difficulties associated with feature selection.

- Development of a genetic algorithm feature selection methodology.

- Analysis and comparisons of the various feature selection methods which will aid in the selection of an appropriate feature selection algorithm.

- A study of Bayesian ANNs, the multiobjective approach, and the various feature selection methods for the computerized detection of mass lesions in mammography.

- A mass CAD systems that has both a simpler and more understandable methodology and a better overall performance.

7 Papers and Presentations

The following papers have been accepted or submitted to peer review journals:

- "Automated Seeded Lesions Segmentation of Digital Mammograms." Matthew A. Kupinski and Maryellen L. Giger. *IEEE Transactions on Medical Imaging*, Vol 17, Iss. 8, Pgs. 510-517, 1998.
- "Feature Selection with Limited Datasets." Matthew A. Kupinski and Maryellen L. Giger. *Medical Physics*, Vol. 26, Iss. 10, Pgs. 2176-2182, 1999. [15]
- "Multiobjective Genetic Optimization of Diagnostic Classifiers with Implications for Generating ROC Curves." Matthew A. Kupinski and Mark A. Anastasio. *IEEE Transactions on Medical Imaging*, Vol. 18, Iss. 8, Pgs. 675-685, 1999. [13]
- "Optimization and FROC Analysis of Rule-Based Detection Schemes Using a Multiobjective Approach." Mark A. Anastasio, Matthew A. Kupinski and Robert M. Nishikawa. *IEEE Transactions on Medical Imaging*, Vol. 17, Iss. 6, Pgs. 1089-1093, 1998. [19]
- "Ideal Observer Estimation Using Bayesian Classification Neural Networks." submitted to *IEEE Transactions on Medical Imaging.*
- "Computerized Detection of Mass Lesions in Mammography Based on Radial Gradient Index." submitted to *Medical Physics.*
- "Classification of Suspect Regions in the Computerized Detection of Mass Lesions in Mammography." submitted to *Medical Physics.*

The following conference proceedings papers have been published:

- "Optimization of Neural Network Inputs with Genetic Algorithms." Digital Mammography '96, Chicago, Illinois. 1996.
- "Feature Selection and Classifiers for the Computerized Detection of Mass Lesions in Digital Mammography." Matthew A. Kupinski and Maryellen L. Giger. IEEE International Congress on Neural Networks, Houston, Texas. 1997.
- "Investigation of Regularized Neural Networks for the Computerized Detection of Mass Lesions in Digital Mammograms." Matthew A. Kupinski and Maryellen L. Giger. IEEE EMBS, Chicago, Illinois, 1997.
- "A Multiobjective Approach to Optimizing Computer-Aided Diagnosis Schemes." M. A. Anastasio, M. A. Kupinski, R. M. Nishikawa, and M. L. Giger. IEEE MIC, Toronto, Canada, 1998.
- "Multiobjective Genetic Optimization of Diagnostic Classifiers Used in the Computerized Detection of Mass Lesions in Mammography." M. A. Kupinski and M. L. Giger, SPIE Medical Imaging Conference, San Diego, California. 2000.

- A Comparison of Bayesian ANN and Multiobjective Classifier Training Using Limited Datasets." CARS, San Francisco, California, 2000.

The following presentations have been given:

- "Feature Selection with Limited Datasets," Matthew A. Kupinski and Maryellen L. Giger, RSNA 1998.
- "Multiobjective Optimization of Diagnostic Classifiers and its Relationship to ROC Analysis and the Ideal Observer," Mark A. Anastasio and Matthew A. Kupinski, Future Directions in Nuclear Medicine Physics and Engineering, Chicago, IL, 1999.
- "Multiobjective Optimization of Diagnostic Classifiers: Pareto Optimality and the Ideal Observer," Mark A. Anastasio and Matthew A. Kupinski, Eighth Far West Image Perception Conference, Alberta, Canada.
- "Ideal Observer Estimation With Bayesian Classification Neural Networks," Matthew A. Kupinski, Darrin C. Edwards, and Maryellen L. Giger, Eighth Far West Image Perception Conference, Alberta, Canada.
- "Bayesian Artificial Neural Networks in the Computerized Detection of Mass Lesions," Matthew A. Kupinski, Darrin C. Edwards, Maryellen L. Giger, and Alexandra E. Baehr, American Association of Physicists in Medicine, Nashville, Tennessee, 1999.
- "Computerized Detection of Mass Lesions in Digital Mammography Using Radial Gradient Index Filtering," Matthew A. Kupinski and Maryellen L. Giger, presented at RSNA 1999.
- "Computerized Detection of Mass Lesions using Feature Filtering," Matthew A. Kupinski and Maryellen L. Giger, Presented at the AAPM World Congress 2000.

8 Conclusions

We have studied some of the fundamental properties of feature selection. We have found that the probability of selecting an optimal subset of features rapidly decreases as the sample size decreases and the total number of features from which to select increases. Understanding the limitation of feature selection will help us select (using methods such as 1D analysis and genetic algorithms) a useful and robust subset of features to be used in the computerized detection of mass lesions in mammography.

We have also studied the use of Bayesian artificial neural networks in classification tasks. We have found that Bayesian ANNs produce more accurate and, yet, robust solutions to classification problems. Bayesian ANNs also train more rapidly than do conventional ANNs using round-robin methodology. This information will be used design more accurate and robust pattern classifiers for the computerized detection of mass lesions in mammography.

We have developed a novel classifier training approach known as the multiobjective approach. This method has the advantage that it directly optimizes the ROC curve of a classifier and, thus, returns the optimal ROC curve that can be obtained using the give classifier on the given training dataset.

We have introduced a new initial filtering scheme to detect mass lesions in mammography. The performance of feature selection methods and of pattern classifiers is limited by the performance of the initial detection algorithm. We have shown that RGI filtering substantially outperformed the previous method of detecting mass lesions known as bilateral subtraction.

Finally, we have applied all of the above techniques to the computerized detection of mass lesions in mammography and produced a CAD system that is both simpler and has a better overall performance over previous techniques.

References

[1] E. Silverberg, C. C. Boring, and T. S. Squires, *Cancer Statistics*, vol. 40. 1990.

[2] L. W. Basset and R. H. Gold, *Breast Cancer Detection. Mammography and Other Methods in Breast Imaging.* Grune and Stratton, 1987.

[3] J. Lissner, M. Kessler, and G. Anhalt, "Developments in methods for early detection of breast cancer," in *Early Breast Cancer* (J. Aandler and J. Baltzer, eds.), (Berlin), Springer-Verlag, 1984.

[4] I. Andersson, "What can we learn from interval carcinomas?," *Recent Results in Cancer Research*, vol. 90, pp. 191–193, 1984.

[5] C. J. Baines, A. B. Miller, and C. Wall, "Sensitivity and specificity of first screen mammography in the Canadian national breast screening study. A preliminary report from five centers," *Radiology*, vol. 160, pp. 295–298, 1986.

[6] J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancers missed by mammography," *American Journal of Roentgenology*, vol. 132, pp. 737–739, 1979.

[7] S. R. Pollei, F. A. Mettler, S. A. Bartow, G. Moradian, and M. Moskowitz, "Occult breast cancer. Prevalence and radiographic detectability," *Radiology*, vol. 16, pp. 459–462, 1987.

[8] J. B. Buchanan, J. S. Spratt, and L. S. Heuser, "Tumor growth, doubling times, and the inability of the radiologist to diagnose certain cancers," *Radiologic Clinics of North America*, vol. 21, pp. 115–126, 1983.

[9] T. Holland, M. Mrvunac, J. H. C. L. Hendricks, and B. Bekker, "So-called interval cancers of the breast. Pathologic and radiographic analysis," *Cancer*, vol. 49, pp. 2527–2533, 1982.

[10] E. L. Thurfjell, K. A. Lernevall, and A. A. Taube, "Benefit of independent double reading in a population-based mammography screening program," *Radiology*, vol. 191, pp. 241–244, 1994.

[11] C. E. Metz and J.-H. Shen, "Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis," *Medical Decision Making*, vol. 12, pp. 60–75, 1992.

[12] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Transactions on Medical Imaging*, 2000 (in review).

[13] M. A. Kupinski and M. A. Anastasio, "Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 675–685, 1999.

[14] M. A. Kupinski and M. L. Giger, "Feature selection and classifiers for the computerized detection of mass lesions in digital mammography," in *Proceedings of the 1997 International Conference on Neural Networks (ICNN '97)*, (Houston, TX), pp. 2460–2463, IEEE/ICNN, June 9-12 1997.

[15] M. A. Kupinski and M. L. Giger, "Feature selection with limited datasets," *Medical Physics*, vol. 26, pp. 2176–2182, 1999.

[16] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 510–517, 1998.

[17] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Investigation of feature-analysis techniques," *Journal of Digital Imaging*, vol. 7, pp. 18–26, 1994.

[18] D. M. Levine, *Users guide to the PGAPack parallel genetic algorithm librar y.* In: Report ANL-95/18 Argonne National Laboratory, 1996.

[19] M. A. Anastasio, M. A. Kupinski, and R. M. Nishikawa, "Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 1089–1093, 1998.

Ideal Observer Approximation Using Bayesian Classification Neural Networks

Matthew A. Kupinski, Darrin C. Edwards,
Maryellen L. Giger, and Charles E. Metz

Department of Radiology, MC2026
The University of Chicago
5841 South Maryland Avenue
Chicago, IL 60637

This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202) and USPHS grants RR11459, T32 CA09649, and GM 57622.

November 18, 1999

Abstract

It is well understood that the optimal classification decision variable is the likelihood ratio or any monotonic transformation of the likelihood ratio. An automated classifier which maps from an input space to one of the likelihood ratio family of decision variables is an optimal classifier or "ideal observer." Artificial neural networks (ANNs) are frequently used as classifiers for many problems. In the limit of large training sample sizes, an ANN approximates a mapping function which is a monotonic transformation of the likelihood ratio, *i.e.*, it estimates an ideal observer decision variable. A principal disadvantage of conventional ANNs is the potential over-parameterization of the mapping function which results in a poor approximation of an optimal mapping function for smaller training samples. Recently, Bayesian methods have been applied to ANNs in order to regularize training to improve the robustness of the classifier. A Bayesian ANN should thus better approximate the optimal decision variable given small sample sizes. We have evaluated the accuracy of Bayesian ANN models of ideal observer decision variables as a function of the number of hidden units used, the signal-to-noise ratio of the data, and the number of features or dimensionality of the data. We show that when enough training data are present, excess hidden units do not substantially degrade the accuracy of Bayesian ANNs. However, the minimum number of hidden units required to best model the optimal mapping function varies with the complexity of the data.

Keywords

Bayesian neural networks, ideal observers, ROC analysis, computer-aided diagnosis

I. Introduction

In image analysis and computer-aided diagnosis, automated classifiers are often employed to determine whether a region within an image is abnormal (with a specified disease) or normal (without that disease) [1–3]. Typically, features are extracted from a suspicious site to form a vector of input features which is projected or mapped onto a scalar decision variable using the parameters of the classifier. A threshold is then applied to this decision variable to determine whether the input features are representative of an abnormal or normal region. Before any classification can be performed, the parameters of the classifier must be determined. Classifier "training" involves using a dataset of observations or features from both the normal class and the abnormal class to determine the parameters of the classifier so it will perform acceptably on future datasets of unknown pathology.

It is well understood that the optimal classification decision variable is the likelihood ratio or any monotonic transformation of the likelihood ratio [4–6]. The ROC curve [4,7–9]

produced using one of the likelihood ratio family of decision variables is the optimal ROC curve which best describes the limiting tradeoffs between sensitivity and specificity. Any system that uses the likelihood ratio or a monotonic transformation of the likelihood ratio to make decisions is known as an ideal observer [10].

A classification artificial neural network (ANN) can be viewed as a highly parameterized classifier or mapping function [11–14]. The output of an ANN in the limit of large sample sizes approximates a mapping function which is a monotonic transformation of the likelihood ratio [12, 15], *i.e.*, it approximates an ideal observer decision variable. A principal disadvantage of conventional ANNs is the possible over-parameterization of the mapping function, which results in a poor approximation of the optimal mapping function given small training sample sizes. Recently, Bayesian methods have been applied to ANNs [16, 17] in order to regularize training to improve the robustness of the classifier. A Bayesian ANN should thus better approximate the optimal decision variable given small sample sizes. Many researchers have analyzed the performance of classification ANNs by evaluating classification ability through ROC analysis on training and testing datasets [18, 19]. Bayesian ANNs, however, are trained not only to produce an ideal observer ROC curve, but also to produce a particular ideal observer mapping function. Hence, an evaluation of a Bayesian ANN's ability to approximate this particular mapping function is of fundamental importance.

In this paper, we investigate the accuracy of Bayesian ANN models of an ideal observer decision variable given simulated datasets of various sizes and neural networks with a single hidden layer. Section II contains a brief introduction to the optimal classifier decision variable, Bayesian ANNs, and the connection between the two. Section III examines a one-dimensional example in detail to provide insights regarding Bayesian ANNs. Section IV describes the methods used for implementing a Bayesian ANN and evaluating its accuracy. In Section V, we study the effects of sample size, input dimension, number of weights, and signal-to-noise ratio of the data on the accuracy of Bayesian ANNs. Finally, Sections VI and VII provide a discussion of the results and a summary of the advantages and disadvantages of Bayesian ANNs for classification.

II. Background

A. Optimal Decision Variables and ROC Analysis

The task of an automated two-group classifier is to determine whether an observation comes from the abnormal class denoted by π_a or the complementary normal class denoted by π_n. The features corresponding to an observation can be expressed as a D-dimensional random vector $\vec{\boldsymbol{x}} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$, where boldface type denotes random variables. Projection classifiers, such as ANNs, classify by mapping an observation onto a new random variable $\boldsymbol{y} = g(\vec{\boldsymbol{x}})$, where $g(\cdot)$ is the mapping function. An observation $\vec{\boldsymbol{x}}$ is classified as abnormal if $g(\vec{\boldsymbol{x}}) \geq y_c$ where y_c is the decision variable threshold. One should note that defining the abnormal assignment as greater than or equal to the threshold y_c is an arbitrary convention. We could, equivalently, have defined the abnormal class as corresponding to decision variable outcomes that are strictly less than, strictly greater than, or less than or equal to y_c but we will, without loss of generality, use the definition given above throughout this paper.

The conditional density functions of data from the abnormal and normal classes are given by $p_{\vec{x}}(\vec{x}|\pi_a)$ and $p_{\vec{x}}(\vec{x}|\pi_n)$ respectively. Similarly, the conditional density functions of the decision variable are given by $p_y(y|\pi_a)$ for the abnormal class and $p_y(y|\pi_n)$ for the normal class. We use the symbol p to denote both continuous and discrete density functions, with its subscripts denoting the random variables drawn from that density function. The true-positive fraction, or the expected fraction of abnormal cases classified correctly, is given by

$$TPF(y_c) = \int_{y_c}^{\infty} p_y(y|\pi_a)\, dy = \int_{\{\vec{x}:g(\vec{x})\geq y_c\}} \cdots \int p_{\vec{x}}(\vec{x}|\pi_a)\, d^D\vec{x}, \tag{1}$$

whereas the false-positive fraction (the expected fraction of normal cases incorrectly classified) is

$$FPF(y_c) = \int_{y_c}^{\infty} p_y(y|\pi_n)\, dy = \int_{\{\vec{x}:g(\vec{x})\geq y_c\}} \cdots \int p_{\vec{x}}(\vec{x}|\pi_n)\, d^D\vec{x}. \tag{2}$$

By varying y_c over its entire range and plotting the $(FPF(y_c), TPF(y_c))$ pairs, one obtains a receiver operating characteristic (ROC) curve [4, 7–9], which describes the performance tradeoffs achievable by the classifier.

It has been shown that the optimal ROC curve is obtained when the mapping function employed is the likelihood ratio

$$\boldsymbol{y} = LR(\vec{\boldsymbol{x}}) \equiv \frac{p_{\vec{x}}(\vec{\boldsymbol{x}}|\pi_a)}{p_{\vec{x}}(\vec{\boldsymbol{x}}|\pi_n)} \tag{3}$$

or any monotonic transformation of the likelihood ratio [20]. Any classifier that does not use the likelihood ratio or a monotonic transformation thereof as its mapping function is suboptimal.

Another commonly employed discriminant function is the posterior probability of an abnormal observation, $p_t(\pi_a|\vec{\boldsymbol{x}})$, called the Bayes optimal discriminant function [5]. Here $\boldsymbol{t}$ is the discrete class membership random variable which takes on the values π_a and π_n. Bayes' rule immediately gives

$$p_t(\pi_a|\vec{x}) = \frac{p_{\vec{x}}(\vec{x}|\pi_a)p_t(\pi_a)}{p_{\vec{x}}(\vec{x}|\pi_a)p_t(\pi_a) + p_{\vec{x}}(\vec{x}|\pi_n)p_t(\pi_n)}. \tag{4}$$

This leads to

$$p_t(\pi_a|\vec{\boldsymbol{x}}) = \frac{k\, LR(\vec{\boldsymbol{x}})}{1 + k\, LR(\vec{\boldsymbol{x}})} \tag{5}$$

where $k \equiv \frac{p_t(\pi_a)}{p_t(\pi_n)}$ and $LR(\vec{\boldsymbol{x}})$ is given in Eqn. 3. Because Eqn. 5 is a monotonic transformation of the likelihood ratio for positive k, we conclude that $p_t(\pi_a|\vec{\boldsymbol{x}})$ is an optimal discriminant function.

B. Modeling and Approximating Probability Functions with ANNs

In statistical estimation, one often makes an assumption concerning the data distribution. For example, the density function of a particular type of data may be modeled as normal with mean μ and standard deviation σ, which we represent as $p_x(x|\mu, \sigma)$. Estimation theory can then be employed along with a sample of data to determine appropriate values of the parameter estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ even when the true density function of the data $p_x(x)$ is not normal. (The circumflex here indicates an estimated quantity.) In this sense the estimation task is analogous to curve fitting or function approximation: the estimated parameters are chosen so that a Gaussian model for the density function "best" approximates the underlying density function from which the sample of data was drawn.

There is potential for confusion in using the vertical bar to denote nonrandom parameters of a function, because this notation is identical to that for a conditional probability;

many authors therefore adopt a different notation for this case (a semicolon instead of a bar, for example). However, in estimation theory it is often useful to consider similar problems from very different points of view, regarding the "parameters" of a function as nonrandom parameters, or as estimates which are functionally dependent upon a set of random observational data, or even as Bayesian quantities derived from density functions (in a Bayesian sense) that are not necessarily dependent on the observational data. Because we wish to emphasize the similarities between these approaches, rather than the subtle (*albeit* fundamental) differences between them, we will use the vertical bar notation to denote (*a*) dependence on nonrandom parameters of a function (such parameters will be clearly stated as being nonrandom); (*b*) conditional dependence on the particular value indicated of a random variable (such variables will be clearly stated as being random); (*c*) conditional dependence on the particular value of an estimate derived from random observational data (distinguished from (*b*) by a circumflex indicating an estimated quantity); and (*d*) a conditional density considered as a function of the random variable (or estimate) upon which it is conditional (distinguished from (*b*) and (*c*), respectively, by boldface type indicating a random variable).

Thus in an expression such as $p_{\boldsymbol{y}}(y|x, \hat{\mu}_x, \vec{w})$, x represents a particular value of the random variable $\boldsymbol{x}$ upon which $\boldsymbol{y}$ depends conditionally (we assume that $\boldsymbol{x}$ was previously introduced as a random variable); $\hat{\mu}_x$ is a particular value of the estimate $\hat{\boldsymbol{\mu}}_x$, upon which $\boldsymbol{y}$ also depends conditionally; and $\vec{w}$ is a set of other (nonrandom) parameters of the conditional density function of $\boldsymbol{y}$.

In the normal example described above, the relevant density function was modeled as a function dependent upon a small number of parameters (namely μ and σ). One may also consider more complicated cases in which the underlying model for the density functions from which observational data are drawn does not have such a simple closed-form representation and may depend upon a large number of parameters $\vec{w}$. The ANN function is one such example. An ANN is a set of connected nodes based loosely on the human neuron system that maps a generally multidimensional input vector $\vec{x}$ to a generally multidimensional output vector $\vec{y}$ using the parameters $\vec{w}$ [11–14,16]. In this paper we will be dealing with ANNs that map $\vec{x}$ to a scalar value y. The output of this ANN function

is given by

$$y = s(\vec{w}^{(n-1)} \cdot \vec{x}^{(n-1)} + w_0^{(n-1)}) \tag{6}$$

where s is a sigmoidal activation function (*e.g.* tanh or the logistic cumulative distribution function), and where $\vec{x}^{(i)}$ represents the output of the ith hidden layer of the ANN, defined recursively as

$$x_j^{(i)} = s(\vec{w}^{(i-1)(j)} \cdot \vec{x}^{(i-1)} + w_0^{(i-1)(j)}), \tag{7}$$

where $\vec{x}^{(o)} \equiv \vec{x}$, the input. Readers unfamiliar with ANNs may consult references [11, 12].

Since an ANN architecture of sufficient complexity is known to be able to closely approximate any continuous function [21], it is reasonable to assume that nearly any underlying density function for $\vec{x}$ can be closely approximated by a function in this family of functions for some value of a sufficiently large set of parameters $\vec{w}$. As explained above, we write such an approximation as $p_{\vec{x}}(\vec{x}|\vec{w})$ to emphasize that this is an approximation to some underlying function $p_{\vec{x}}(\vec{x})$.

C. Maximum Likelihood Estimation of ANN Weights

Since $p_t(\pi_a|\vec{x})$ in Eqn. 5 is an optimal discriminant function, we can treat the task of classification as one of modeling, or approximating, the function $p_t(\pi_a|\vec{x})$. As outlined in the previous section, we approximate the optimal discriminant function by an ANN function, writing it as $p_t(\pi_a|\vec{x}, \vec{w})$, the range of which is bound between 0 and 1. We then define $p_t(\pi_n|\vec{x}, \vec{w}) \equiv 1 - p(\pi_a|\vec{x}, \vec{w})$. The practical task, given an actual ANN with a finite number of weights and a sample of training data, is to choose weights $\hat{\vec{w}}$ such that the difference between the ANN output $p_t(\pi_a|\vec{x}, \hat{\vec{w}})$ and the true Bayes optimal discriminant function $p_t(\pi_a|\vec{x})$ is small.

ANN training involves using a training dataset $\{\boldsymbol{X}, \boldsymbol{T}\}$ to determine a value of $\hat{\vec{w}}$, where $\boldsymbol{X} = \{\vec{x}_i\}_{i=1}^N$ is the set of training feature vectors, $\boldsymbol{T} = \{t_i\}_{i=1}^N$ is the set of known truth (π_a or π_n) for each feature vector, and N is the total number of samples in the training dataset. Assuming the data to be independently sampled, the true likelihood of the data is given by

$$p_{X,T}(X, T) = \prod_{i=1}^{N} p_t(t_i|\vec{x}_i) p_{\vec{x}}(\vec{x}_i). \tag{8}$$

If $p_t(\pi_a|\vec{x})$ is approximated by an ANN function $p_t(\pi_a|\vec{x},\vec{w})$, then the likelihood of the data assuming the ANN model and given a particular value of the parameters $\vec{w}$ is

$$p_{X,T}(X,T|\vec{w}) = \prod_{i=1}^{N} p_t(t_i|\vec{x}_i,\vec{w})p_{\vec{x}}(\vec{x}_i). \tag{9}$$

Maximizing the likelihood of the data, $p_{X,T}(X,T|\vec{w})$, with respect to $\vec{w}$ is equivalent to maximizing $p_T(T|X,\vec{w})$ with respect to $\vec{w}$. The maximum likelihood estimate $\hat{\vec{w}}^{ML}$ is thus obtained by maximizing

$$p_T(T|X,\vec{w}) = \prod_{i=1}^{N} p_t(t_i|\vec{x}_i,\vec{w}) \tag{10}$$

with respect to $\vec{w}$. The logarithm of the above expression is known as the cross-entropy error function [12].

Ruck *et al.* [15] proved that an ANN trained using the sum of squares error function approaches the Bayes optimal discriminant function in the limit of infinite training data. We wish to show that maximizing the utility function in Eqn. 10 also yields a discriminant function that approaches the Bayes optimal discriminant function given infinite training data. Taking the logarithm of Eqn. 10 yields

$$\ln(p_T(T|X,\vec{w})) = \sum_{i=1}^{N} \ln(p_t(t_i|\vec{x}_i,\vec{w})). \tag{11}$$

Taking the limit as N approaches infinity and employing the Strong Law of Large Numbers [22], we arrive at

$$\ln(p_T(T|X,\vec{w})) = \int\cdots\int [\ln(p_t(\pi_a|\vec{x},\vec{w}))p_{\vec{x}}(\vec{x},\pi_a) + \ln(p_t(\pi_n|\vec{x},\vec{w}))p_{\vec{x}}(\vec{x},\pi_n)]\, d^D\vec{x}. \tag{12}$$

Since a sufficiently large ANN is known to be able to closely approximate any continuous function [21], it is reasonable to assume that $p_t(\pi_a|\vec{x},\vec{w})$ can take on any functional form of interest here. The task of finding the $\vec{w}$ that maximizes Eqn. 12 is replaced with the more tractable problem of finding the function $p_t(\pi_a|\vec{x},\vec{w})$ that maximizes a particular *functional*, namely the integral on the right side of Eqn. 12.

Using the calculus of variations [23], one can show that Eqn. 12 is maximized when $p_t(\pi_a|\vec{x},\vec{w}) = p_t(\pi_a|\vec{x})$. It follows that $p_{X,T}(X,T|\vec{w}) = \prod_{i=1}^{N} p_t(t_i|\vec{x}_i,\vec{w})p_{\vec{x}}(\vec{x}_i)$ is maximized *in the limit of infinite data* by the $\hat{\vec{w}}^{ML}$ such that $p_t(\pi_a|\vec{x},\hat{\vec{w}}^{ML}) = p_t(\pi_a|\vec{x})$.

This result is of limited practical utility because one does not have infinite data from which to estimate $\hat{\vec{w}}^{ML}$. In practice, when given a sample of data $\{\boldsymbol{X}, \boldsymbol{T}\}$ from the population, the ML ANN model attempts to approximate the *sample* Bayes optimal discriminant function, which is one everywhere an abnormal observation is located in the training dataset, zero everywhere a normal observation is located in the training dataset, and arbitrary (or undefined) elsewhere. This discontinuous behavior is one of the drawbacks of the maximum likelihood method that the Bayesian methods attempt to address.

D. Bayesian Estimation of ANN Weights

We have argued that the maximum likelihood method approximates the *sample* Bayes optimal discriminant function when one has finite training data. It is necessary, however, to approximate the *population* Bayes optimal discriminant function, because the sample Bayes optimal discriminant function is of little practical use. This is conventionally performed by regularization methods such as early stopping [11,24] and weight decay [13,25,26]. Recently, Bayesian methods have been used to regularize the training process to approximate the population Bayes optimal discriminant function rather than the sample discriminant function [16,17]. Training a Bayesian ANN involves choosing the maximum *a posteriori* (MAP) weight vector $\hat{\vec{w}}^{MAP}$ which maximizes

$$p_{\vec{w}}(\vec{w}|X,T,\vec{\alpha}) = \frac{p_{X,T}(X,T|\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})}{\int p_{X,T}(X,T|\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})d\vec{w}}, \tag{13}$$

where $p_{X,T}(X,T|\vec{w})$ is given in Eqn. 9, and the "prior" $p_{\vec{w}}(\vec{w}|\vec{\alpha})$ is the regularization term that employs the parameters $\vec{\alpha}$ to incorporate our prior belief of what constitute reasonable values of $\vec{w}$. It should be noted that $p_{X,T}(X,T|\vec{w})$ in Eqn. 13 is now a true conditional probability in the Bayesian sense [27], whereas $\vec{w}$ was a nonrandom parameter of the underlying model in Eqn. 9. The term $p_{X,T}(X,T|\vec{w})$ also does not depend on $\vec{\alpha}$, because we specifically model this function as depending only on the weight vector $\vec{w}$. The prior is often modeled as

$$p_{\vec{w}}(\vec{w}|\vec{\alpha}) = C\exp\left(-\frac{1}{2}\sum_{i=1}^{W}\alpha_i w_i^2\right) \tag{14}$$

where α_i is the regularization constant for the ith weight w_i, C is a constant ensuring that $p_{\vec{w}}(\vec{w}|\vec{\alpha})$ is a properly normalized density function, and W is the number of elements in the weight vector $\vec{w}$ [12,16,17]. Equation 14 attempts to limit the magnitude of the weight

vector $\vec{w}$, because it has been shown that large weight values correspond to complicated mapping functions [16, 28].

The denominator in Eqn. 13, referred to hereafter as the evidence, is often written as

$$p_{X,T}(X,T|\vec{\alpha}) \equiv \int p_{X,T}(X,T|\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})d\vec{w}. \tag{15}$$

We are faced with the same problem that motivated Eqn. 13, that is, choosing appropriate values of $\vec{\alpha}$. We again adopt a Bayesian view of the relevant parameters and use Bayes' rule to arrive at

$$p_{\vec{\alpha}}(\vec{\alpha}|X,T) = \frac{p_{X,T}(X,T|\vec{\alpha})p_{\vec{\alpha}}(\vec{\alpha})}{p_{X,T}(X,T)}. \tag{16}$$

If we assume that $p_{\vec{\alpha}}(\vec{\alpha})$ is a flat prior, then maximizing $p_{\vec{\alpha}}(\vec{\alpha}|X,T)$ with respect to $\vec{\alpha}$ is equivalent to maximizing $p_{X,T}(X,T|\vec{\alpha})$, or the evidence, with respect to $\vec{\alpha}$. In principle, one could use the data to first determine the paramters $\vec{\alpha}$ using Eqn. 15, and then determine the parameters of the neural network $\vec{w}$ using Eqn. 13. In practice, however, it is prohibitively time-consuming to evaluate the integral in Eqn. 15, so approximations of the integral are often employed [16]. For a more detailed discussion of Bayesian ANNs and the methods used to determine the regularization constants $\vec{\alpha}$, see references [16] and [17].

As previously stated, Bayesian methods are useful primarily when one has finite training data. It is still important, however, for the Bayesian method to arrive at the Bayes optimal discriminant function in the limit of infinite training data in order for the estimator to be consistent. We wish to show that maximizing the right side of Eqn. 13 given infinite data again leads to the Bayes optimal discriminant function. Note that the denominator of Eqn. 13 does not depend on $\vec{w}$, whereas the numerator is the same as Eqn. 9 but with an extra factor $p_{\vec{w}}(\vec{w})$. Taking the logarithm of Eqn. 13, we arrive at

$$\ln(p_{\vec{w}}(\vec{w}|X,T,\vec{\alpha})) = \sum_{i=1}^{N} \ln(p_t(t_i|\vec{x}_i,\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})^{1/N}) + \sum_{i=1}^{N} \ln(p_{\vec{x}}(\vec{x}_i)) - \ln\left(\int p_{X,T}(X,T|\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})d^W\vec{w}\right). \tag{17}$$

Note that the maximization of Eqn. 17 with respect to $\vec{w}$ affects only the first term on the right side and that, with respect to $\vec{x}_i$ and t_i, this first term is similar in form to the right side of Eqn. 11. The remaining arguments from the previous section are unchanged, and

it can readily be shown that the same results are obtained, namely that $p_{\vec{w}}(\vec{w}|X,T,\vec{\alpha})$ is maximized *in the limit of infinite data* by the $\hat{\vec{w}}^{MAP}$ such that $p_t(\pi_a|\vec{x},\hat{\vec{w}}^{MAP}) = p_t(\pi_a|\vec{x})$.

As in the maximum likelihood situation, this result is of little practical importance because one never has infinite data. We also know that the maximum likelihood method performs poorly with "small" training datasets because the maximum likelihood method attempts to approximate the *sample* Bayes optimal discriminant function instead of the population Bayes optimal discriminant function. It is not known, however, how well the Bayesian method performs with limited training data. In this work we determine empirically the ability of Bayesian ANNs to approximate the Bayes optimal discriminant function when trained using finite datasets.

III. One-Dimensional Study

Before proceeding to the methods and results of the experiments, we will present a one-dimensional example of the behavior of a Bayesian ANN. This will allow us to visually present and explain the Bayesian ANN at a greater level of detail than possible in the higher dimensional experiments. Our goal in this section is to provide a framework for the interpretation of the methodology and results which will be presented in the next two sections.

Given the 1D density functions defined by

$$p_x(x|\pi_n) = \left(\frac{1}{2\pi}\right)^{1/2} \exp\left(-\frac{x^2}{2}\right) \tag{18}$$

and

$$p_x(x|\pi_a) = \left(\frac{b^2}{2\pi}\right)^{1/2} \exp\left(-\frac{(bx-a)^2}{2}\right) \tag{19}$$

with a and $b > 0$, we can compute the theoretical optimal mapping function $p_t(\pi_a|x)$ using Eqn. 5, which results in

$$y = p_t(\pi_a|x) = \frac{kb \exp\left(-\frac{1}{2}((bx-a)^2 - x^2)\right)}{1 + kb \exp\left(-\frac{1}{2}((bx-a)^2 - x^2)\right)}. \tag{20}$$

Using the theory of random variables [29], we can determine the conditional density functions of the new random variable $\boldsymbol{y} = p_t(\pi_a|\boldsymbol{x})$ using

$$p_y(y|t) = \frac{p_x(x_{(1)}|t)}{|p_t'(\pi_a|x_{(1)})|} + \cdots + \frac{p_x(x_{(l)}|t)}{|p_t'(\pi_a|x_{(l)})|} + \cdots, \tag{21}$$

where t is either π_a or π_n, $x_{(l)}$ denotes the lth root of Eqn. 20 for a specific y, and $p'_t(\pi_a|x_{(l)})$ is the derivative of $p_t(\pi_a|x)$ with respect to x evaluated at $x_{(l)}$ [29]. For the special case $b = 1$ and $a > 0$, Eqn. 20 has only one root given by

$$x_{(1)} = \frac{1}{a}\log\left(\frac{y}{k(1-y)}\right) + \frac{a}{2}. \tag{22}$$

Combining Eqns. 20, 21, and 22, we obtain the $b = 1$ conditional density functions of y,

$$p_y(y|\pi_n) = \frac{1}{ay(1-y)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{1}{a}\log\left(\frac{y}{k(1-y)}\right) + \frac{a}{2}\right)^2\right) \tag{23}$$

and

$$p_y(y|\pi_a) = \frac{1}{ay(1-y)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{1}{a}\log\left(\frac{y}{k(1-y)}\right) - \frac{a}{2}\right)^2\right), \tag{24}$$

for $0 < y < 1$.

The $b \neq 1$ case is more complicated because Eqn. 20 now has two roots

$$x_{(1)} = \frac{ba}{b^2-1} - \sqrt{\frac{a^2}{(b^2-1)^2} - \frac{2}{b^2-1}\log\left(\frac{y}{kb(1-y)}\right)} \tag{25}$$

and

$$x_{(2)} = \frac{ba}{b^2-1} + \sqrt{\frac{a^2}{(b^2-1)^2} - \frac{2}{b^2-1}\log\left(\frac{y}{kb(1-y)}\right)}. \tag{26}$$

Using Eqns. 25, 26 and 20 with 21 we obtain the $b \neq 1$ conditional density functions of y

$$p_y(y|\pi_n) = \frac{p_x(x_{(1)}|\pi_n) + p_x(x_{(2)}|\pi_n)}{y(1-y)\sqrt{a^2 - 2(b^2-1)\log(\frac{y}{kb(1-y)})}} \tag{27}$$

and

$$p_y(y|\pi_a) = \frac{p_x(x_{(1)}|\pi_a) + p_x(x_{(2)}|\pi_a)}{y(1-y)\sqrt{a^2 - 2(b^2-1)\log(\frac{y}{kb(1-y)})}}. \tag{28}$$

The values of y are bound between 0 and 1 due to the form of Eqn. 3. Because the square root terms in Eqns. 25 and 26 must be non-negative, the value of b further restricts the range of possible y (or $p(\pi_a|x)$) values when $b \neq 1$, *i.e.*,[1]

$$b > 1 \quad \rightsquigarrow \quad 0 < y < \left(1 - \frac{1}{kb\,\exp\left(\frac{a^2}{2(b^2-1)}\right) + 1}\right), \tag{29}$$

[1]The symbol $\rightsquigarrow$ is known as the "leads to" symbol.

$$b < 1 \quad \rightsquigarrow \quad \left(1 - \frac{1}{kb \exp\left(\frac{a^2}{2(b^2-1)}\right) + 1}\right) < y < 1, \tag{30}$$

and

$$b = 1 \quad \rightsquigarrow \quad 0 < y < 1. \tag{31}$$

Given a sample of 1D data taken from the densities shown in Eqns. 37 and 38, the Bayesian ANN should model the mapping function given in Eqn. 20. To show this, we sampled 500 normal observations and 500 abnormal observations ($N = 1000$) from the density functions shown in Fig. 1 ($a = 1$ and $b = 0.5$). We trained a Bayesian ANN with 10 hidden units on this data to produce $p_t(\pi_a|x, \hat{\vec{w}}^{MAP})$, which is shown in Fig. 2 along with the theoretical true mapping function $p_t(\pi_a|x)$ (Eqn. 20). (The training method will be described in the next section.) Because the Bayesian ANN mapping function approximates the true mapping function $p_t(\pi_a|x)$, we conclude that the output $y_i = p_t(\pi_a|x_i, \hat{\vec{w}}^{MAP}); i = 1, \ldots, N$ of the Bayesian ANN using the training data x_i as input should behave as if sampled from the density functions given in Eqns. 25-28. Figure 3 shows both the theoretical density functions $p_y(y|\pi_n)$ and $p_y(y|\pi_a)$ for the $a = 1$, $b = 0.5$ case, together with the histograms of the abnormal and normal training data output from the Bayesian ANN. The histograms of the Bayesian ANN decision variable closely match the densities of the ideal observer decision variable. Finally, one can generate the theoretical ROC curve using Eqns. 1 and 2 and the theoretical densities of the ideal decision variable $p_y(y|\pi_a)$ and $p_y(y|\pi_n)$. This can then be compared (see Fig. 4) with the ROC curve generated using the Bayesian ANN output of an independently sampled testing dataset using the "proper" curve fitting method [6, 30]. Again, the true optimal ROC curve and the ROC curve produced using the output of the Bayesian ANN are similar.

IV. Methods

A. Bayesian ANN Implementation

In this work, we have employed a neural network with an input layer, a single hidden layer, and a single output node. Our implementation of the Bayesian ANN is based on the work of MacKay [16]. An approximation is made in MacKay's methods that should be noted. In order to determine the regularization parameters $\vec{\alpha}$, one must evaluate the

integral in Eqn. 15. In MacKay's method, the integrand $p_{X,T}(X,T|\vec{w})p_{\vec{w}}(\vec{w}|\vec{\alpha})$ is locally approximated using a Taylor expansion around the most probable weight values $\vec{\boldsymbol{w}}^{MP}$ as being Gaussian. The integral in Eqn. 15 can then be evaluated explicitly. This changes the method of optimization, however, because the evidence now depends on this $\vec{\boldsymbol{w}}^{MP}$, which in turn depends upon the choice of $\vec{\alpha}$. Hence, a dual optimization is performed in which one fixes the parameters $\vec{\alpha}$ to find $\vec{\boldsymbol{w}}^{MP}$, then uses $\vec{\boldsymbol{w}}^{MP}$ to find a new set of parameters $\vec{\boldsymbol{\alpha}}$, and then repeats this process a fixed number of times. At this point, the parameters $\vec{\boldsymbol{\alpha}}$ are assumed to be properly determined and the $\vec{\boldsymbol{w}}^{MP}$ can be labeled the maximum *a posteriori* value of the weights $\vec{\boldsymbol{w}}^{MAP}$. It has been determined empirically that eight such iterations work well for most applications. In other methods, such as those discussed by Neal [17], the integral in Eqn. 15 is evaluated using Monte Carlo methods. This paper will not evaluate the validity of the assumption made in using MacKay's methods, instead focusing on the accuracy of Bayesian ANNs under these assumptions.

The optimizations were performed using variable metric (or quasi-Newton) techniques [31] with a tolerance of 1.0×10^{-7} and a maximum of 1000 iterations. Numerical methods were employed to determine the covariance matrix of the Gaussian approximation used to evaluate the evidence [12]. In practice, the constants $\vec{\alpha}$ are constrained so that only three distinct values are used: one for the hidden layer weights, one for hidden layer biases, and one for output layer weights and biases. Reference [16] provides details concerning Bayesian ANNs and the implementation we used.

B. ROC Analysis with Bayesian ANNs

Given a Bayesian ANN mapping function

$$\boldsymbol{y} = p_t(\pi_a|\vec{\boldsymbol{x}}, \hat{\vec{\boldsymbol{w}}}^{MAP}) \tag{32}$$

and the density functions $p_y(y|\pi_n)$ and $p_y(y|\pi_a)$, one could produce the ROC curve for this classifier using Eqns. 1 and 2. It is interesting to note that a Bayesian ANN approximates an optimal mapping function without explicitly modeling any of the density functions $p_{\vec{x}}(\vec{x}|\pi_n)$, $p_{\vec{x}}(\vec{x}|\pi_a)$, $p_y(y|\pi_n)$ or $p_y(y|\pi_a)$. We can therefore use a Bayesian ANN to model the optimal mapping function from feature space $\vec{\boldsymbol{x}}$ to decision space $\boldsymbol{y}$, but we cannot employ a Bayesian ANN to produce a continuous estimate of the optimal ROC curve.

Metz and Pan [6, 30] developed a parametric ROC curve fitting model that assumes an underlying binormal model and uses likelihood ratio. The ROC curves produced here were generated using the maximum likelihood estimates of the distribution parameters under this "proper" binormal model given testing dataset outputs of the Bayesian ANN.

C. Methods of Evaluation

ROC analysis cannot directly assess the performance of Bayesian ANNs in the task of approximating $p_t(\pi_a|\vec{\boldsymbol{x}})$, because monotonic transformations of $p_t(\pi_a|\vec{\boldsymbol{x}})$ will produce identical ROC curves. Figure 5 shows two different mapping functions $p_t(\pi_a|x)$ and $p_t(\pi_a|x, \hat{\vec{w}}^{MAP})$ and their corresponding ROC curves. Because these two mapping functions are monotonic transformations of one another, the ROC curves produced by these mapping functions are identical. In this work, however, we take the stance that the task of a Bayesian ANN is not only to generate the ideal observer ROC curve, but also to approximate accurately the *particular* ideal observer decision variable $p_t(\pi_a|\vec{\boldsymbol{x}})$. Although the ideal observer ROC will be obtained if a Bayesian ANN produces $p_t(\pi_a|\vec{\boldsymbol{x}})$ exactly, the primary task in training Bayesian ANNs is not to optimize with respect to the ROC curve but with respect to the underlying decision variable. We therefore believe that an analysis of the Bayesian ANN's ability to approximate the specific decision variable $p_t(\pi_a|\vec{\boldsymbol{x}})$ is of fundamental importance. Other methods of training that directly acknowledge the multiobjective nature of classifier training can be employed to directly optimize with respect to the ROC curve and not with respect to an underlying decision variable [20, 32].

As we have discussed, a Bayesian ANN produces an approximation $p_t(\pi_a|\vec{\boldsymbol{x}}, \hat{\vec{\boldsymbol{w}}}^{MAP})$ of $p_t(\pi_a|\vec{\boldsymbol{x}})$. If the densities $p_{\vec{x}}(\vec{x}|\pi_a)$ and $p_{\vec{x}}(\vec{x}|\pi_n)$ are known, we can compute the theoretical $p_t(\pi_a|\vec{\boldsymbol{x}})$ using Eqn. 5. The mean-squared-error (MSE) between the true optimal mapping function $p_t(\pi_a|\vec{\boldsymbol{x}})$ and the Bayesian ANN model of this function using one particular sample of training data $S = \{X, T\}$ is defined as

$$\epsilon(S) = E_{\vec{x}}\{(p_t(\pi_a|\vec{\boldsymbol{x}}, \hat{\vec{w}}^{MAP}) - p_t(\pi_a|\vec{\boldsymbol{x}}))^2\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\vec{x}}(\vec{x}) \left[p_t(\pi_a|\vec{x}, \hat{\vec{w}}^{MAP}) - p_t(\pi_a|\vec{x})\right]^2 d^D\vec{x}. \quad (33)$$

Here $\hat{\vec{w}}^{MAP}$ is not random because it is the specific set of weights determined from the

data S. This expectation value becomes difficult to compute as the dimensionality of the input data increases, so we estimate $\epsilon(S)$ from a testing dataset $\boldsymbol{S}' = \{\boldsymbol{X}', \boldsymbol{T}'\}$ with N' observations:

$$\hat{\boldsymbol{\epsilon}}(S, \boldsymbol{S}') = \frac{1}{N'} \sum_{i=1}^{N'} \left[p_t(\pi_a | \vec{\boldsymbol{x}}_i', \hat{\vec{w}}^{MAP}) - p_t(\pi_a | \vec{\boldsymbol{x}}_i') \right]^2 \tag{34}$$

where $\vec{\boldsymbol{x}}_i'$ is the ith observation in the testing dataset $\boldsymbol{S}'$. Equations 33 and 34 measure the difference between the true optimal mapping function $p_t(\pi_a|\vec{x})$ and one particular Bayesian ANN model of this function weighted by the density of the data $p_{\vec{x}}(\vec{x})$. We wish to measure not only the difference between a single Bayesian ANN model and the true optimal mapping function, but also the robustness of the methodology. Therefore we average multiple observations of $\hat{\boldsymbol{\epsilon}}(\boldsymbol{S}, \boldsymbol{S}')$ where $\boldsymbol{S}$ is now random, *i.e.*,

$$\hat{\boldsymbol{\mu}} = \frac{1}{M} \sum_{i=1}^{M} \hat{\boldsymbol{\epsilon}}(\boldsymbol{S}_i, \boldsymbol{S}_i'). \tag{35}$$

Here $\boldsymbol{S}_i$ and $\boldsymbol{S}_i'$ are distinct and independently sampled pairs of training and testing datasets. It is important to note that Eqn. 35 measures the average MSE over M different Bayesian ANN models with M different estimates of $\hat{\vec{\boldsymbol{w}}}^{MAP}$. Similarly, an estimate of the variance of $\hat{\boldsymbol{\mu}}$, the standard error squared, is given by

$$\hat{\boldsymbol{s}^2} = \frac{1}{M(M-1)} \sum_{i=1}^{M} (\hat{\boldsymbol{\epsilon}}(\boldsymbol{S}_i, \boldsymbol{S}_i') - \hat{\boldsymbol{\mu}})^2. \tag{36}$$

For all the studies we performed, the number (M) of different datasets used to estimate the sample mean and standard error was fixed at 100.

D. Simulations

In all simulation studies reported in this paper, we sampled data using an isotropic Gaussian

$$p_{\vec{x}}(\vec{x}|\pi_n) = \left(\frac{1}{2\pi}\right)^{D/2} \exp\left(-\sum_{i=1}^{D} \frac{x_i^2}{2}\right) \tag{37}$$

with zero mean and unit marginal standard deviations for the normal class and an isotropic Gaussian

$$p_{\vec{x}}(\vec{x}|\pi_a) = \left(\frac{b^2}{2\pi}\right)^{D/2} \exp\left(-\sum_{i=1}^{D} \frac{(bx_i - a)^2}{2}\right) \tag{38}$$

with marginal means a/b and marginal standard deviations of $1/b$ for each feature for the abnormal class. Here D is the dimension of $\vec{x}$ or, equivalently, the number of features used.

As proposed and derived in reference [18], the signal-to-noise ratio (SNR) for this data is defined such that the "signal" is the difference between the means of the distributions and the "noise" is the root-mean-square standard deviation of the two distributions. This results in

$$\mathrm{SNR}^{(D)} = \sqrt{\frac{2D}{b^2+1}}\, a = \sqrt{D}\ \mathrm{SNR}^{(1)} \tag{39}$$

where $\mathrm{SNR}^{(1)}$ is the signal-to-noise ratio of one-dimensional data with parameters a and b.

We studied the accuracy of Bayesian ANN models as a function of the number of training samples N, the number of hidden units employed H, the SNR of the data, and the dimension of the data D. The relationship between the number of hidden units H and the total number of weights W for a single-output classification neural network is given by $W = (D+2)H + 1$.

V. Results

Figure 6 shows the performance of the Bayesian ANN with varying numbers of hidden units H and input dimensions D for a fixed signal-to-noise ratio $SNR = 1.26$ and sample size $N = 200$. For the lower dimensional curves ($D = 1$, $D = 2$, and $D = 3$), the average MSE between the optimal decision variable and the Bayesian ANN approximation of that decision variable decreases as the number of hidden units increases and then becomes relatively constant after a certain threshold. For $D = 2$, the difference becomes relatively constant after 3 hidden units, whereas the $D = 3$ curve flattens out after 4 hidden units. More parameters are required to better approximate the optimal mapping function as the number of dimensions increases. The $D = 4$ and $D = 5$ curves show a definite minimum in their MSE. The "curse of dimensionality" is a common problem caused by the relative density of data decreasing substantially as the dimensionality of the data increases. This effect is seen in the $D = 4$ and $D = 5$ curves. The Bayesian ANN does not have enough data to properly approximate the optimal mapping function, so a tradeoff exists between simpler solutions with few parameters that cannot match the ideal mapping function (due to under-parameterization), and more complicated solutions with many hidden units that cannot be properly determined due to the lack of data. This effect causes the minimum

in the $D = 4$ and $D = 5$ curves.

Figure 7 shows a similar plot with varying numbers of hidden units H and input dimension D with a fixed $SNR = 1.26$ but with more training samples, $N = 1000$. The Bayesian ANN no longer has any difficulty with the $D = 4$ and $D = 5$ cases; they both flatten out after a certain number of hidden units. In this plot, we can also more clearly see the effect of the dimensionality of the data on the number of hidden units required. The higher the input dimension, the more weights are required to best approximate the optimal decision variable.

The complexity of the optimal mapping function is, for our simulation studies, a function of the SNR of the data. In general, SNR by itself is not sufficient to characterize the relationship between data density and the complexity of the optimal mapping function. This relationship for one-dimensional data was explored by Metz and Pan [6] but is beyond the scope of this work. For our purposes it is sufficient to note that, for a fixed $b \neq 1$, the optimal mapping function (Eqn. 20) becomes more complicated where $p_{\vec{x}}(\vec{x})$ is nonnegligible as SNR decreases, and becomes more sigmoidal where $p_{\vec{x}}(\vec{x})$ is nonnegligible as SNR increases (see Fig. 8). We therefore conclude that the number of weights needed to best model the optimal mapping function should change with SNR for a fixed b. This is shown in Fig. 9 where the SNR is fixed at the larger value of 3.80. Fewer weights (free parameters) are needed to model the optimal mapping function. In fact, increasing the number of hidden units H results in a gradual increase in the average MSE for the $D = 5$ curve. This increase is, however, inconsequential when compared to the average MSE values caused by having too few hidden units in Fig. 7.

In order to further understand the effect of SNR on the accuracy of Bayesian ANNs, we performed simulation studies in which the SNR was varied between 0 and 5. Figure 10 shows the average MSE of Bayesian ANNs as a function of SNR for a fixed $H = 4$ and $N = 200$. The average MSE remains relatively constant and then decreases as SNR increases for the $D = 1$, $D = 2$ and $D = 3$ curves. The remainder of the curves in Fig. 10 show a decrease in the average MSE as the SNR increases. Figure 11 shows the effect of SNR on the accuracy of Bayesian ANNs with $H = 10$ and $N = 1000$. All of the curves in Fig. 11 remain relatively constant until the average MSE begins to decrease as the SNR

increases. These two figures show that the Bayesian ANN can easily model an optimal mapping function that is sigmoidal for values of $\vec{x}$ where $p_{\vec{x}}(\vec{x})$ is nonnegligible (high SNR and $b = 0.5$), but has more difficulty modeling a mapping function that is more complicated where $p_{\vec{x}}(\vec{x})$ is nonnegligible (low SNR and $b = 0.5$). Another method of describing this result is to note that for a high SNR and $b = 0.5$, the quadratic separation function through the data can be approximated by a line where the data are dense, whereas for low SNR and $b = 0.5$, the separation function has much more local curvature where the data are dense.

We have shown the effect of the number of hidden units H, the SNR of the data, and the dimension of the data D on the accuracy of Bayesian ANNs. We have also indirectly shown the effect of sample size on the accuracy ability of Bayesian ANNs because Figs. 6 and 7 have different sample sizes N. Figure 12 more clearly shows the effect of sample size on the average MSE between the true optimal mapping function and the Bayesian ANN model of that function with various numbers of hidden units. As the number of training samples increases, the average MSE decreases. When there are too few hidden units, the average MSE asymptotically approaches a high value but flattens out after just 100 samples. When enough enough hidden units are used to properly estimate the optimal mapping function (see Fig. 7), the average MSE asymptotically approaches a much lower value, but requires more samples to do so. The effect of adding additional hidden units beyond the minimum required to properly model the optimal mapping function has negligible effect.

VI. Discussion

While the average MSE between the optimal mapping function and the Bayesian approximations of that function is fundamentally important, it does not address all the issues relevant to training. If the average MSE is high, one does not know whether the Bayesian ANN mapping function is too complicated, too simple, or, possibly, a monotonic transformation of the ideal observer mapping function. Typically, when too few hidden units are employed in the Bayesian ANN, the resulting mapping function is overly simple, resulting in a large average MSE. When too many hidden units are used without enough training data (as for the $D = 5$ curve in Fig. 6), we have found that the Bayesian ANN

mapping function can become too complicated, *i.e.*, the Bayesian ANN overtrains. Figure 13 shows the training and testing dataset ROC curves for two different Bayesian ANNs. Figure 13(a) was generated with $D = 5$, $N = 160$, $H = 10$, and $SNR = 1.26$. The training dataset ROC curve is clearly well above the ideal observer ROC curve, while the ROC curve generated using an independent testing dataset is well below the ideal observer ROC curve. The Bayesian ANN overtrained in this example. When more samples were used ($N = 1000$), the training and testing dataset ROC curves were similar to the ideal observer ROC curve (Fig. 13(b)).

As we have previously argued, it is important that a Bayesian ANN not only approximate an ideal observer decision variable, but that it approximate the particular ideal observer decision variable $p_t(\pi_a|\vec{x})$. Because the Bayesian ANN models a probability function, it has the added benefit of allowing a specific interpretation of the ANN output. From decision theory [11], we know that one should call an observation $\vec{x}$ abnormal if

$$[U(\pi_a|\pi_a) - U(\pi_n|\pi_a)]p_t(\pi_a|\vec{x}) > [U(\pi_n|\pi_n) - U(\pi_a|\pi_n)](1 - p_t(\pi_a|\vec{x})) \qquad (40)$$

where $U(i|j)$ is the utility of classifying an observation as i when it is actually from class j. Therefore, if one can quantify the utilities associated with classifying observations, then the Bayesian ANN output should be employed to optimally determine a threshold at which to call an observation abnormal. If the output of an ANN does not have this interpretation then one could use the slope of the estimated ROC curve (which is the likelihood ratio over the decision variable) to get an estimate of $p_t(\pi_a|y)$ which could, in turn, be used instead of $p_t(\pi_a|\vec{x})$ in Eqn. 40. This requires an extra estimation step, namely the estimation of the ROC curve, and is, therefore, less direct.

The Bayesian ANN output can also be used to provide probabilities of malignancy to radiologists in observer studies or computerized diagnosis schemes. However, this is difficult with diagnostic classifiers because the prior probabilities are typically very different. The traditional method of sampling is to blindly employ observations, along with their corresponding classes, in training a Bayesian ANN. If the prior probabilities of the two classes are very different, then one class will not be as well represented as the other class. For example, if one randomly chooses 1000 mammograms to be used in ANN training, then there will be, on average, only about 5 cases with malignant lesions. Typically, for

diagnostic classifiers, one samples each class separately so as to get equal numbers of observations in each class. This would cause a Bayesian ANN trained on such data to improperly approximate $p_t(\pi_a|\vec{x})$, because the training data are not properly sampled from the screening population. The Bayesian ANN would, in fact, approximate the function

$$\frac{k'LR(\vec{x})}{1+k'LR(\vec{x})}, \tag{41}$$

where k' is the ratio of malignant to non-malignant training data. We can, however, transform the Bayesian ANN output to have a more appropriate k value. For example, if one samples 500 malignant observations from a population and then samples 500 non-malignant observations from another population, the k' value in Eqn. 41 is 1. If the ratio of the prior probabilities is known to be $k = 0.01$, then one can transform the Bayesian ANN output to approximate $p_t(\pi_a|\vec{x})$ by solving Eqn. 41 for $LR(\vec{x})$ and substituting this function into Eqn. 5 using the appropriate value of k.

We have chosen to analyze Bayesian ANNs that employ a Gaussian prior on the weight values with parameters $\vec{\alpha}$ because such priors have been used extensively in the neural network community. However, other, possibly more appropriate, priors have been developed and studied. For example, Williams [33] has studied the use of a Laplace prior for ANNs, whereas entropy-based and other types of priors are discussed by Buntine and Weigend [34].

VII. Conclusions

We have shown that the goal of training a Bayesian ANN is to approximate a particular ideal observer decision variable. The ROC curves produced using two different decision variables will be the same if the two decision variables are monotonically related. A poor estimate of $p_t(\pi_a|\vec{x})$ can yield an ROC curve equal to that obtained from a good estimate of $p_t(\pi_a|\vec{x})$ if the two solutions are monotonically related. Thus, a measure of a Bayesian ANN's accuracy in approximating the particular ideal observer decision variable $p_t(\pi_a|\vec{x})$ is vital. Because the output of a Bayesian ANN is an estimate of a probability, it has other benefits such as aiding in the determination of decision thresholds and presenting likelihoods of malignancy to radiologists.

As was expected, we have shown that Bayesian ANNs better approximate the optimal

decision variable $p_t(\pi_a|\vec{x})$ when the training dataset has more observations and/or fewer dimensions. We have also shown that, given enough training data, the performance of a Bayesian ANN is not impaired by excess hidden units, as seen in Fig. 7. Above a certain number of hidden units, the average MSE remains relatively constant despite many more parameters being added to the model. With less training data, however (see the $D = 5$ curve in Fig. 6), there is a tradeoff between simplifying the mapping function with fewer hidden units and adding more parameters to allow the mapping function more freedom.

Figure 7 also shows that a minimum number of hidden units is required to "best" approximate the optimal mapping function, because having too few hidden units does not allow the mapping function enough freedom to adequately approximate the optimal mapping function. This is due to the tradeoff between the ANN not having sufficient parameters to effectively approximate the Bayes optimal discriminant function and the ANN having too many parameters with too little training data to effectively estimate a "good" $\hat{\vec{w}}$. Fewer than 6 hidden units for the $D = 5$ curve will result in a poorer approximation of $p_t(\pi_a|\vec{x})$. However, this minimum number of hidden units is dependent on the density functions of the underlying data. When we increased the SNR of the data and, hence, simplified the optimal mapping function, the minimum number of necessary hidden units decreased to 1 (Fig. 9). Given an abundance of training data, it is thus safer to err on the side of more hidden units than fewer, because excess weights are clearly less detrimental than too few. This is not true when one has limited training data.

There is clearly a complicated relationship between the performance of a Bayesian ANN and the number of weights H, the sample size N, the number of input features D, and the signal-to-noise ratio of the data SNR. Care must be taken with each new neural network model to ensure that the optimal decision variable $p_t(\pi_a|\vec{x})$ is being appropriately approximated. In practice, one does not know the true optimal mapping function, so independent test datasets or cross-validation techniques must be used to ensure the robustness of Bayesian ANN classifiers.

ACKNOWLEDGMENTS

The authors thank Mark A. Anastasio and Rufus H. Nagel for their many helpful suggestions and stimulating discussions. This work was supported in parts by grants from the US

Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202) and USPHS grants RR11459, T32 CA09649, and GM 57622.

M. L. Giger and C. E. Metz are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

REFERENCES

[1] M. L. Giger, "Computer-aided diagnosis," *RSNA Categorical Course in Physics*, pp. 283–298, 1993.

[2] K. Doi, M. L. Giger, R. M. Nishikawa, K. R. Hoffmann, H. MacMahon, R. A. Schmidt, and K.-G. Chua, "Digital radiography: A useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images," *Acta Radiologica*, vol. 34, pp. 426–439, 1993.

[3] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer-aided detection of clustered microcalcifications on digital mammograms," *Medical and Biological Engineering and Computing*, vol. 33, pp. 174–178, 1995.

[4] J. Egan, *Signal Detection Theory and ROC Analysis.* Academic Press, New York, 1975.

[5] H. L. Van Trees, *Detection, estimation, and modulation theory (Part I).* New York: Academic Press, 1968.

[6] C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *Journal of Mathematical Psychology*, vol. 43, pp. 1–33, 1999.

[7] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, pp. 283–298, 1978.

[8] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720–733, 1986.

[9] J. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, 1979.

[10] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality: III. ROC metrics, ideal observers, and likelihood-generating functions," *Journal of the Optical Society of America A*, vol. 15, no. 6, pp. 1520–1535, 1998.

[11] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks.* Addison-Wesley, 1989.

[12] C. Bishop, *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press, 1995.

[13] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation.* Redwood City, CA: Addison-Wesley Publishing Company, 1991.

[14] S. Haykin, *Neural Networks, A Comprehensive Foundation.* New York, NY: Macmillan, 1994.

[15] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.

[16] D. J. S. MacKay, *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, Pasadena, California, 1992.

[17] R. M. Neal, *Bayesian Learning for Neural Networks.* Lecture Notes in Statistics, New York: Springer-Verlag Inc., 1996.

[18] Y. Wu, K. Doi, C. E. Metz, N. Asada, and M. L. Giger, "Simulation studies of data classification by artificial

neural networks: Potential applications in medical imaging and decision making," *Journal of Digital Imaging*, vol. 6, pp. 117–125, 1993.

[19] K. Woods and K. W. Bowyer, "Generating ROC curves for artificial neural networks," *IEEE Transactions on Medical Imaging*, vol. 16, no. 3, pp. 329–337, 1997.

[20] M. A. Anastasio and M. A. Kupinski, "Multiobjective optimization of pattern classifiers: Pareto optimality and the ideal observer," tech. rep., The University of Chicago, Chicago, IL, 1999.

[21] K. Hornik, M. Stinchcombe, and H. White, "Multi-layer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[22] S. Ross, *A First Course in Probability*. New York: Macmillan, 1976.

[23] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1963.

[24] W. S. Sarle, "Stopped training and other remedis for overfitting," in *Proceedings of the 27th Symposium on the Interface*, 1995.

[25] C. E. Rasmussen, "Generalization in neural networks," Master's thesis, Technical University of Denmark, August 1993.

[26] D. Rumelhart, "Learning and generalization," in *IEEE International Conference on Neural Networks*, (San Diego), IEEE, 1988.

[27] E. T. Jaynes, "Bayesian methods: General background," in *Maximum Entropy and Bayesian Methods in Applied Statistics* (J. H. Justice, ed.), pp. 1–25, Cambridge University Press, 1986.

[28] M. A. Kupinski and M. L. Giger, "Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms," in *Proceedings of the 19th International Conference of Engineering in Medicine and Biology*, (Chicago, IL), pp. 1336–1339, IEEE/EMBS, Oct. 30-Nov. 2 1997.

[29] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., 1991.

[30] X. Pan and C. E. Metz, "The 'proper' binormal model: Parametric ROC curve estimation with degenerate data," *Academic Radiology*, vol. 4, pp. 380–389, 1997.

[31] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge Unversity Press, 1988.

[32] M. A. Kupinski and M. A. Anastasio, "Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 675–685, 1999.

[33] P. M. Williams, "Bayesian regularization and pruning using a laplace prior," *Neural Computation*, vol. 7, no. 1, pp. 117–143, 1995.

[34] W. L. Buntine and A. S. Weigend, "Bayesian back-propagation," *Complex Systems*, vol. 5, pp. 603–643, 1991.

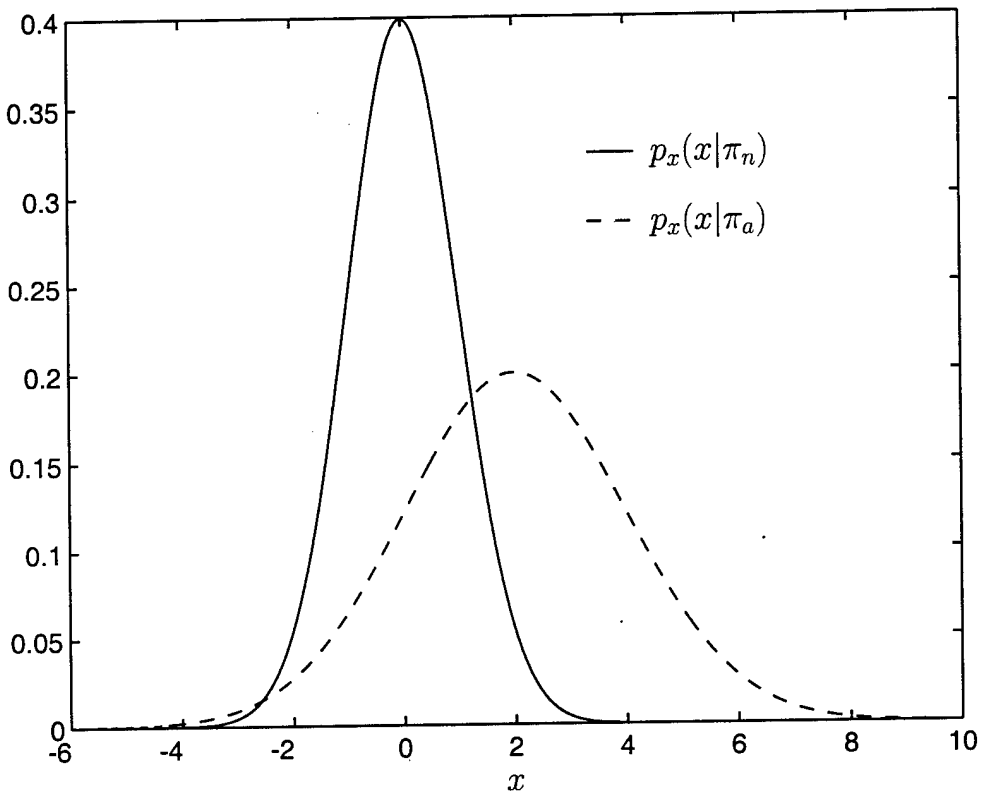


Fig. 1. Example density functions with $a = 1$ and $b = 0.5$. A Bayesian ANN was trained using 500 samples from $p_x(x|\pi_a)$ and 500 samples from $p_x(x|\pi_n)$ to illustrate the estimation task of the Bayesian ANN.

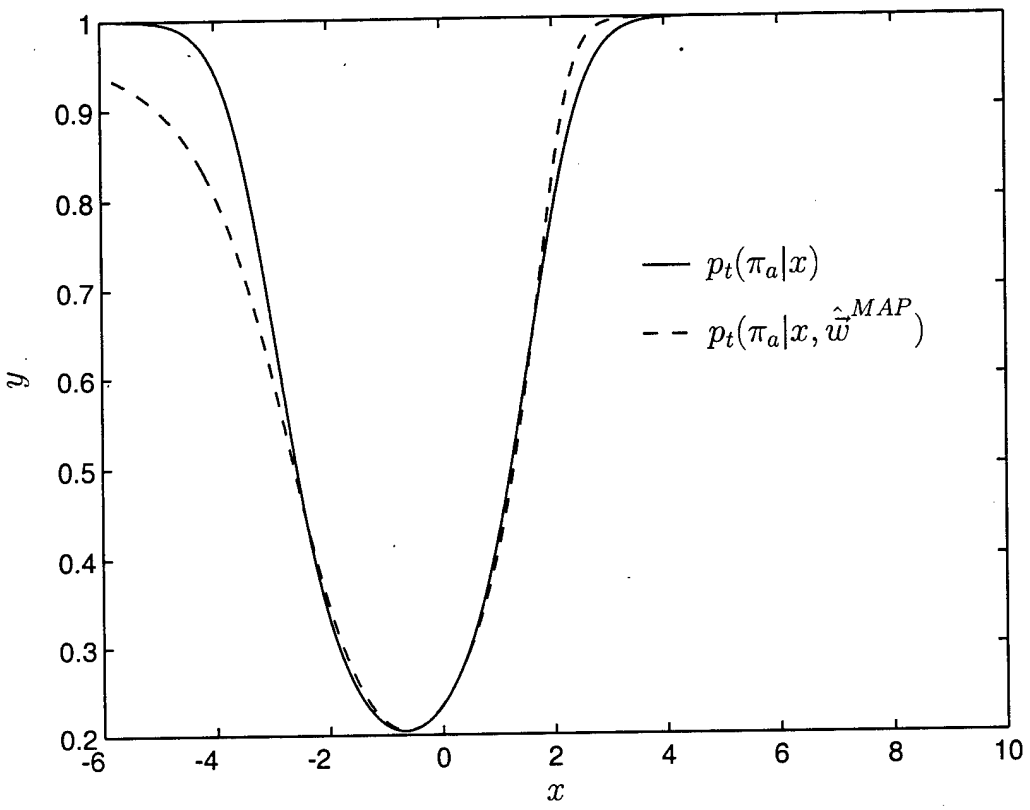
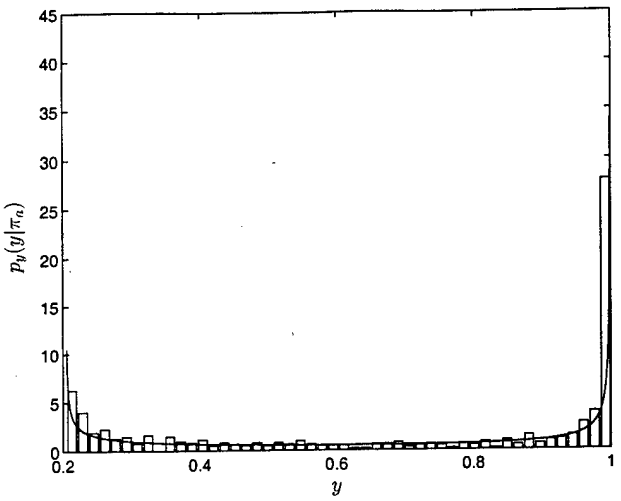
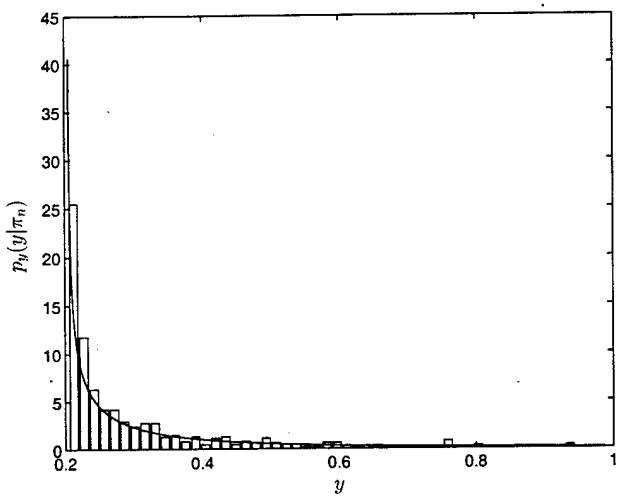


Fig. 2. Using the data sampled from the density functions shown in Fig. 1, the Bayesian ANN produced the mapping function $p_t(\pi_a|x, \hat{\vec{w}}^{MAP})$ (dashed curve) shown with the theoretical mapping function (solid curve) produced using the actual density functions and not the data sampled from the density functions. It should be noted that the largest discrepancies are in the $x < -2$ region of the plot which is an area where there is little data (see densities in Fig. 1). The error calculations employed in this work are weighted by the density of the data $p_x(x)$, so discrepancies in sparse regions of the mapping function are not as important as discrepancies in dense regions.



(a)



(b)

Fig. 3. Given the theoretical mapping function (Eqn. 20) and the original density functions (Eqns. 37 and 38) one can produce the density functions of the decision variable y which is shown (curves) for (a) the abnormal class and (b) the normal class. Plotted with each density function is a histogram of the (a) abnormal training data and (b) the normal training data after it has been mapped using the transfer function $p_t(\pi_a|x, \hat{\vec{w}}^{MAP})$ in Fig. 2. The histograms were scaled to have an area of 1. A k of 1 was used to produce these plots.

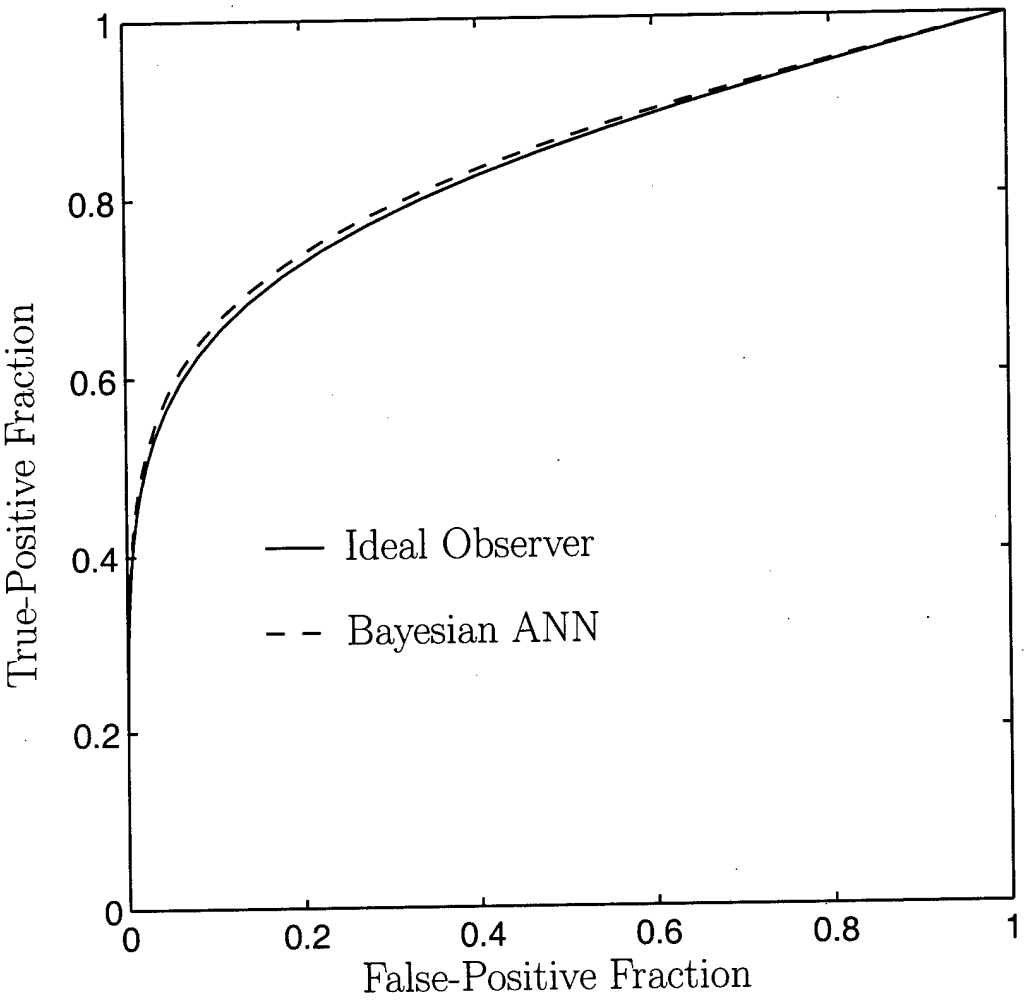
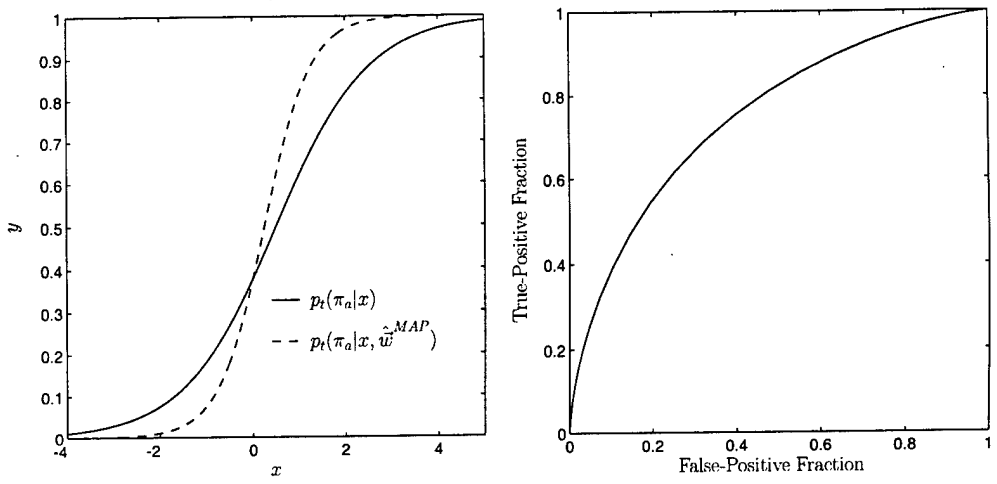


Fig. 4. The ideal observer ROC curve was produced using Eqns. 1 and 2 and the theoretical density functions over the decision variable. The Bayesian ANN ROC curve was produced using the "proper" binormal ROC curve fitting method on the independent testing dataset output y_i of the Bayesian ANN with 1000 samples in each class, *i.e.*, $N = 2000$.



(a) (b)

Fig. 5. Two different mapping functions (a) will produce the same ROC curves (b) if the two mapping functions are monotonic transformations of one another. The Bayesian ANN that generated $p_t(\pi_a|x, \hat{\vec{w}}^{MAP})$ is said to be a poorly trained ANN because it does not approximate $p_t(\pi_a|x)$ well despite the fact that it yields the same ROC curve.

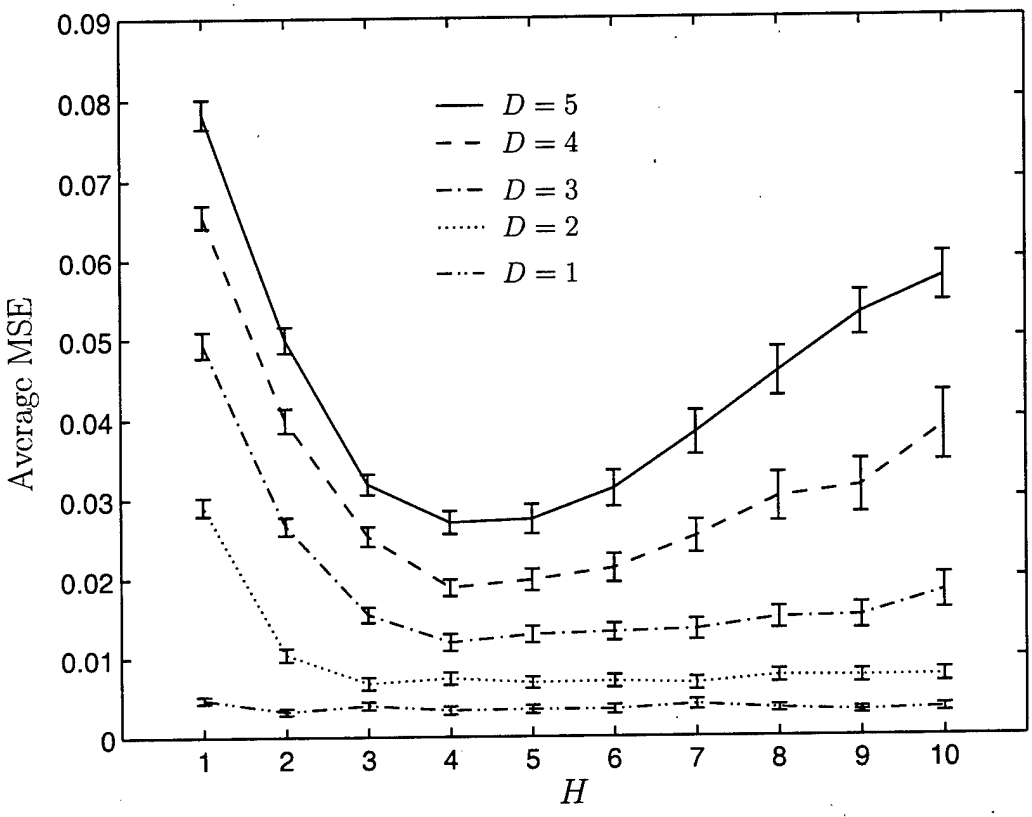


Fig. 6. The effect of the number of hidden units H on the accuracy of Bayesian ANNs with a fixed $SNR = 1.26$ and sample size $N = 200$. Because there is a limited training dataset, the Bayesian ANN cannot properly approximate the optimal mapping function at higher dimensions ($D = 4$ and $D = 5$). The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had parameters $a = 1/\sqrt{D}$ and $b = 0.5$.

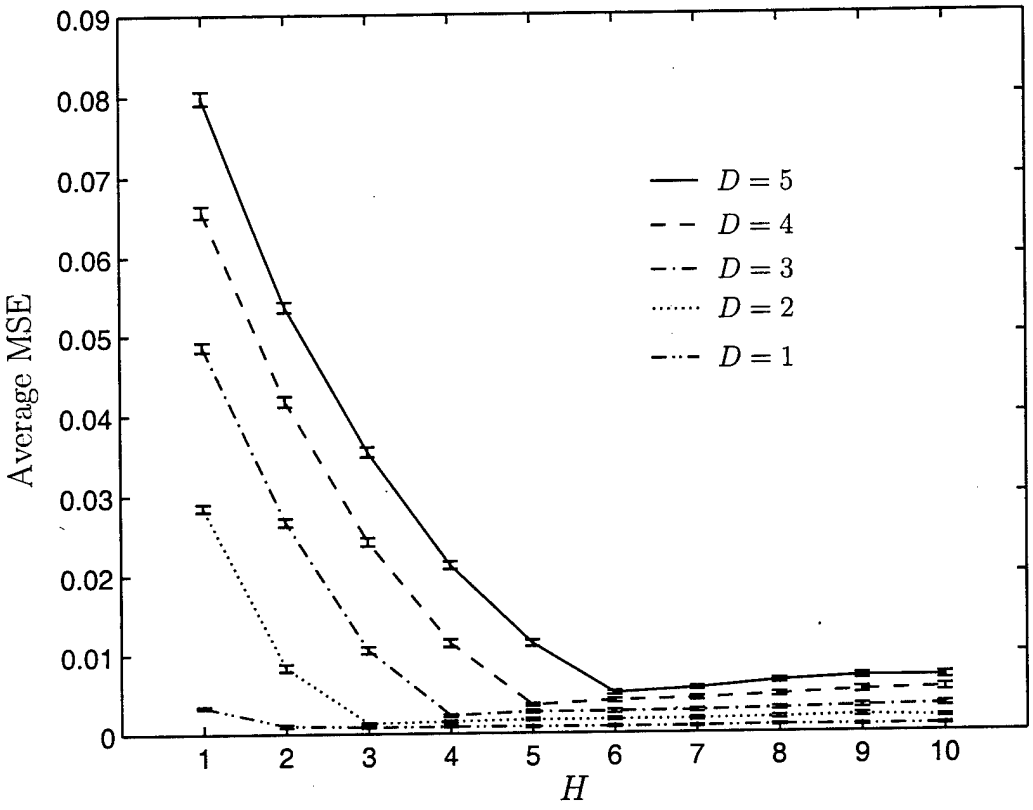


Fig. 7. The effect of the number of hidden units H on the accuracy of Bayesian ANNs with a fixed $SNR = 1.26$ and sample size $N = 1000$. With enough training data, the Bayesian ANN can properly approximate the optimal mapping function at the higher dimensions ($D = 4$ and $D = 5$). The effect of increasing the number of hidden units after a certain point is no longer beneficial but also not very detrimental. The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had parameters $a = 1/\sqrt{D}$ and $b = 0.5$.

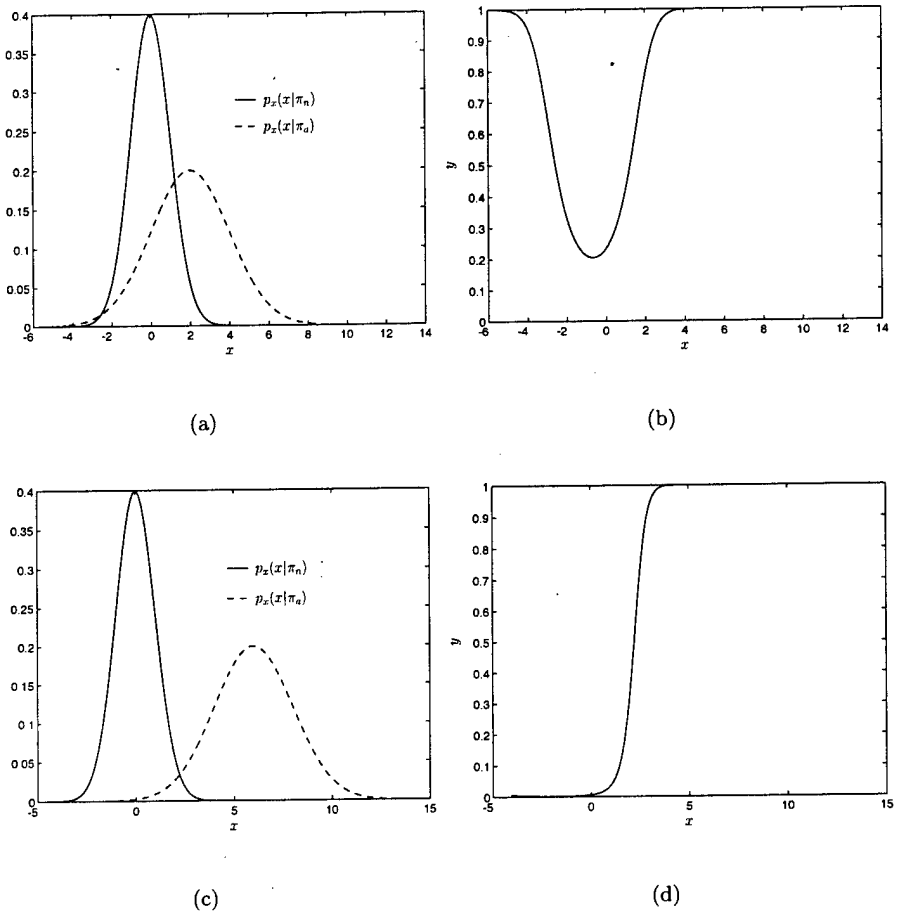


Fig. 8. For (a) low SNR and $b = 0.5$, the (b) optimal mapping function is complicated for values of $\vec{x}$ such that $p_{\vec{x}}(\vec{x})$ is nonnegligible. When (c) the SNR is higher for the same $b = 0.5$, the (d) optimal mapping function is sigmoidal where $p_{\vec{x}}(\vec{x})$ is nonnegligible. Equation 20 always has two roots when $b \neq 1$ so, in actuality, the mapping function shown in (d) does have two values of x for a given y (except for the minimum) but the second x (for most y) occurs in an uninteresting region, *i.e.*, an observation that is unlikely to occur. Analogous conclusions can be made for D larger than 1.

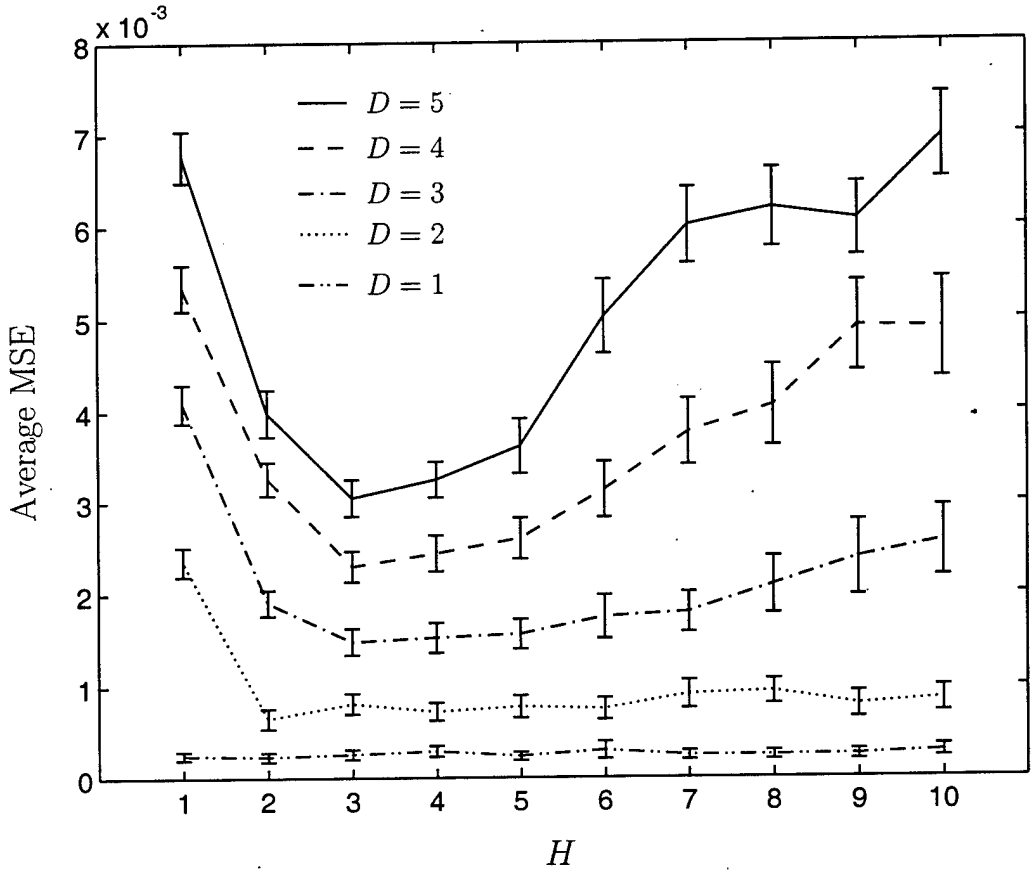


Fig. 9. The effect of the number of hidden units H on the accuracy of Bayesian ANNs with a fixed $SNR = 3.80$ and sample size $N = 1000$. The SNR is high, so the optimal mapping function is sigmoidal in shape where $p_{\vec{x}}(\vec{x})$ is nonnegligible; few hidden nodes are needed to model this optimal mapping function. Note that despite the increase in the average MSE as the hidden units increases, the magnitude of this difference is still small when compared to Fig. 7. The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had parameters $a = 4/\sqrt{D}$ and $b = 0.5$.

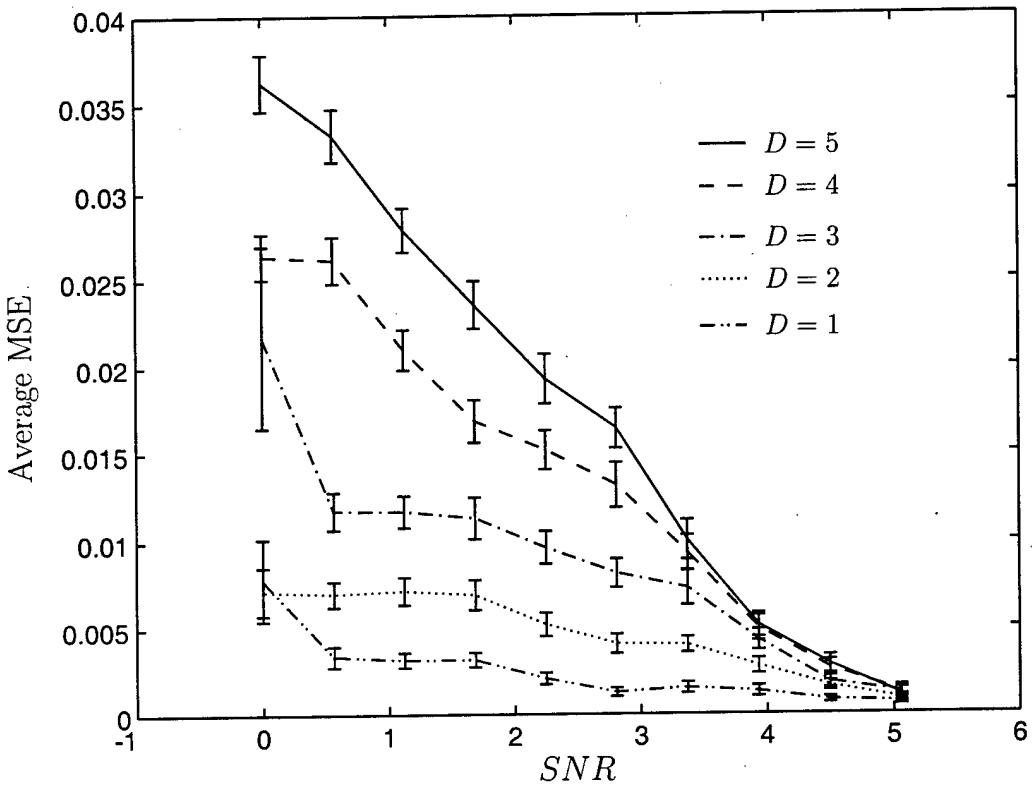


Fig. 10. The effect of the signal-to-noise ratio SNR on the accuracy of Bayesian ANNs with fixed $H = 4$ and $N = 200$. Increasing SNR causes a decrease in the average MSE between the optimal mapping function and the Bayesian ANN model of that function. The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had a fixed $b = 0.5$.

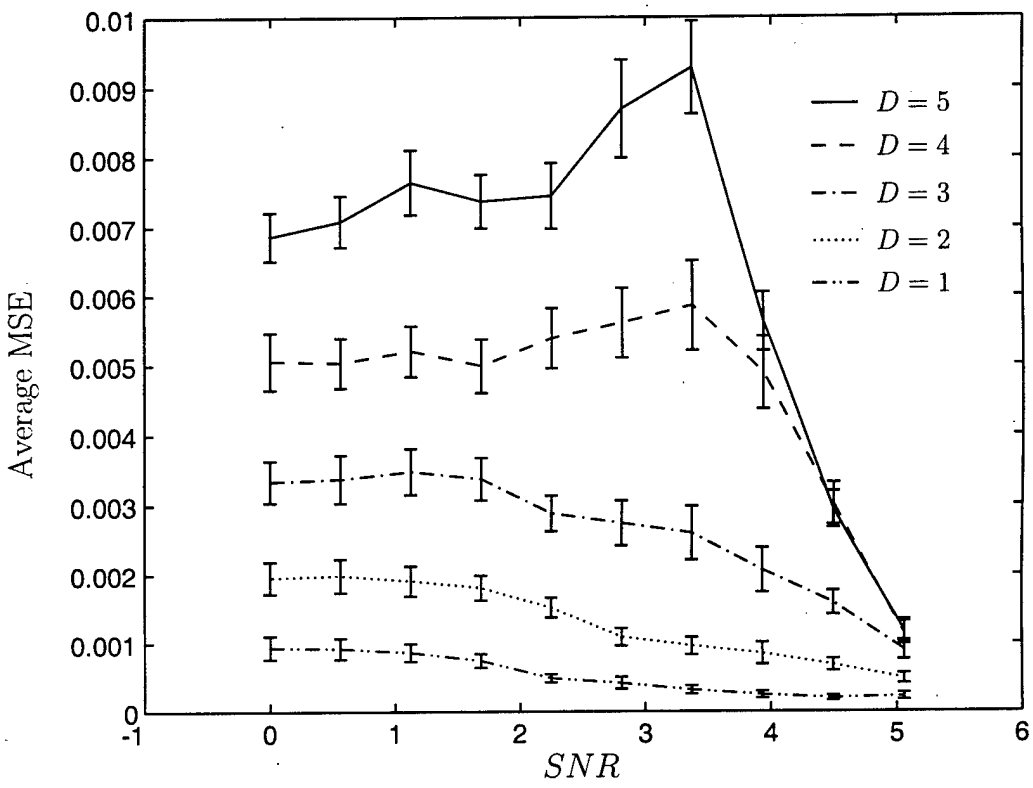


Fig. 11. The effect of the signal-to-noise ratio SNR on the accuracy of Bayesian ANNs with fixed $H = 10$ and $N = 1000$. The average MSE remains relatively constant as SNR increases but falls after a certain point. The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had a fixed $b = 0.5$.

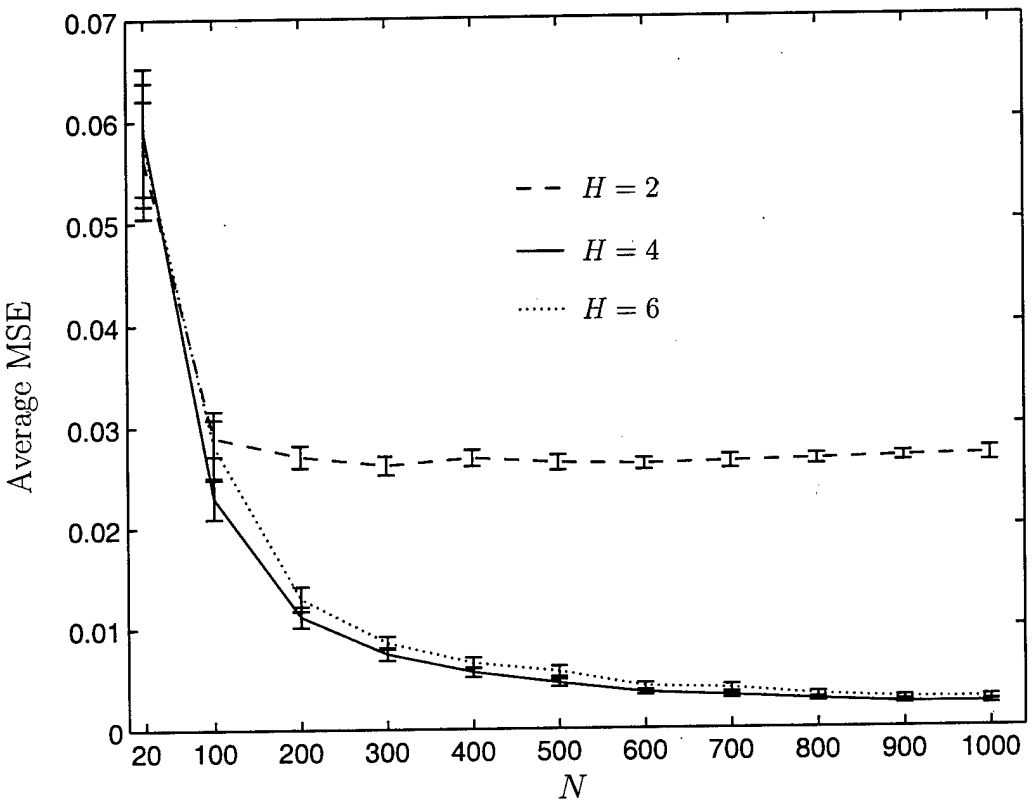


Fig. 12. The effect of sample size N on the accuracy of Bayesian ANNs with $D = 3$, $SNR = 1.26$ and varying H. Increasing the training dataset size results in a better approximation of $p_t(\pi_a|\vec{x})$. With fewer hidden units ($H = 2$), the average MSE flattens out quickly as N increases. When more hidden units are used, the curves ($H = 4$ and $H = 6$) asymptotically approach a much smaller average MSE. The error bars represent ± 2 standard errors of each mean ($\hat{s}$). The density functions used in this simulation study had a fixed $a = 1/\sqrt{D}$ and $b = 0.5$.

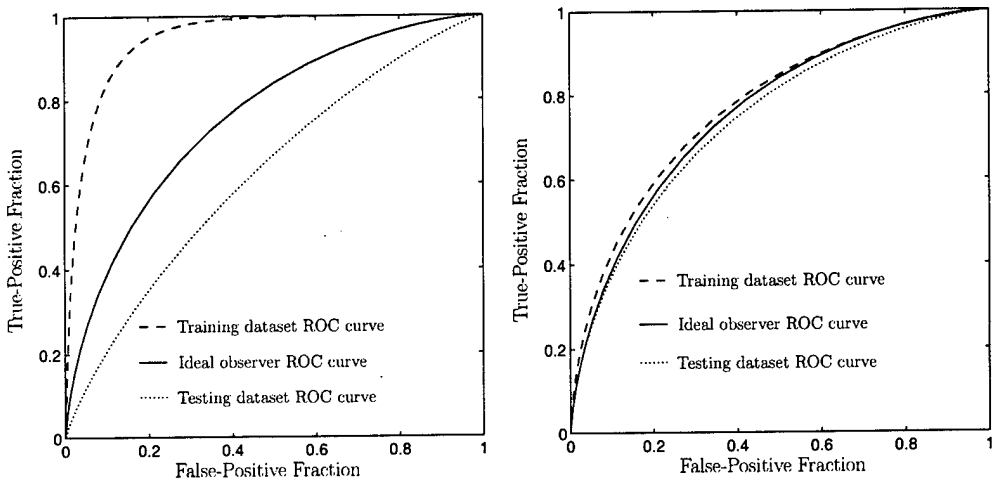


Fig. 13. The training and testing dataset ROC curves generated using $D = 5$, $H = 10$, $SNR = 1.26$ and (a) $N = 160$ or (b) $N = 1000$. When little training data is used with many hidden units, the Bayesian ANN tends to overtrain, resulting in a large training dataset ROC curve and a much lower testing dataset ROC curve. When more samples are used, the training and testing dataset ROC curves are similar to the ideal observer ROC curve. The density functions used in this simulation study had a fixed $a = 1/\sqrt{D}$ and $b = 0.5$. The average MSE between the optimal mapping function and the Bayesian ANN estimates of that function were 0.103 for (a) and 0.005 for (b).

Multiobjective Genetic Optimization of Diagnostic Classifiers with Implications for Generating Receiver Operating Characteristic Curves

Matthew A. Kupinski, *Student Member, IEEE* and Mark A. Anastasio, *Student Member, IEEE*

Abstract—**It is well understood that binary classifiers have two implicit objective functions (sensitivity and specificity) describing their performance. Traditional methods of classifier training attempt to combine these two objective functions (or two analogous class performance measures) into one, so that conventional scalar optimization techniques can be utilized. This involves incorporating *a priori* information into the aggregation method so that the resulting performance of the classifier is satisfactory for the task at hand. We have investigated the use of a niched Pareto multiobjective genetic algorithm for classifier optimization. With niched Pareto genetic algorithms, an objective vector is optimized instead of a scalar function, eliminating the need to aggregate classification objective functions. The niched Pareto genetic algorithm returns a set of optimal solutions that are equivalent in the absence of any information regarding the preferences of the objectives. The *a priori* knowledge that was used for aggregating the objective functions in conventional classifier training can instead be applied post-optimization to select from one of the series of solutions returned from the multiobjective genetic optimization. We have applied this technique to train a linear classifier and an artificial neural network using simulated datasets. The performances of the solutions returned from the multiobjective genetic optimization represent a series of optimal (sensitivity, specificity) pairs, which can be thought of as operating points on an ROC curve. All possible ROC curves for a given dataset and classifier are less than or equal to the ROC curve generated by the niched Pareto genetic optimization.**

Keywords— **Multiobjective optimization, genetic algorithms, diagnostic classifiers, ROC analysis.**

I. Introduction

THE task in medical diagnostic decision making is typically one of employing multiple features to classify an observation as normal or abnormal. A radiologist may, for example, note the size, shape and margin sharpness of a potential breast lesion in a mammogram and somehow use this information to determine whether a cancer is present. In computer-aided diagnosis (CAD) [1–3], computers take features extracted from medical images and determine whether pathology is present by using automated classifiers [4, 5]. It is well known that the optimal method for classifying would be to use the likelihood ratio or any monotonic transformation of the likelihood ratio as the discriminant function [4]. The goal in training a diagnostic classifier is to employ a limited dataset to determine the parameters of the classifier such that it approximates the likelihood ratio decision rule. For the most part, these classifiers work in a similar fashion. A dataset of features extracted from both normal (without disease) and abnormal (with disease) images is used for determining the classifier parameter values, or for "training" the classifier, so that it correctly classifies future datasets of unknown pathology.

Classifier training can be viewed as an optimization problem where the quantity to be maximized is the performance of the classifier on an independent dataset. There are, however, numerous problems with representing classifier performance by a single (scalar) objective function, which is needed so that one can use a scalar optimizer [6, 7]. Binary classifiers [4] have, in essence, two implicit objective functions: one describing how well they classify the abnormal cases (sensitivity) and one describing how well they classify the normal cases (specificity). These two objective functions are non-commensurable, implying that it may not be possible to simultaneously improve both the sensitivity and specificity. Traditional methods of classifier training attempt to combine these two objective functions (or two analogous class performance measures) into a single scalar objective function that permits the use of conventional (scalar) optimization techniques [8]. A drawback of this approach is that the proper way of aggregating the objective functions is usually unknown. There are, in fact, an infinite number of ways of mapping two objective functions to a single scalar function. Even when *a priori* information about the relative importance of the two objective functions is available, it is not always clear how to incorporate it in the aggregating approach to objective function design. Sometimes numerous *ad hoc* combination functions are tried until a suitable objective function is found [8]. Most classifiers do not aggregate sensitivity and specificity directly. Artificial neural networks, for example, typically employ a sum-of-squares error function [5] which can still be thought of as a sum of two non-commensurable objectives, i.e., one objective is to map abnormal observations to a value close to 1 and the other objective is to map normal observations to a value close to 0.

Genetic algorithms (GAs) [9] have been applied to many diagnostic and classification problems [8,10–15]. A conventional GA, however, is a scalar optimization technique. It thus has the undesirable features of an aggregating-based approach. One method of avoiding this is to adopt a multiobjective approach [16,17] to the optimization problem. In a multiobjective optimization approach, the objective function is vector-valued and the independent objectives (sensitivity and specificity) are optimized simultaneously. Thus,

This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202) and USPHS grants CA24806 and RR11459.

the need to aggregate the independent objective functions is removed. Unlike a scalar optimization that returns a single solution, the solution to the multiobjective optimization problem is a set of solutions called the Pareto-optimal set. The Pareto-optimal set is defined as the set of solutions for which no other solution exists that is better in both objectives. In the absence of any preference information about the objectives, the members of the Pareto-optimal set are equally valid solutions to the optimization problem; no other solutions exist that are better in all of the objectives. In the context of diagnostic classifier optimization, the members of the Pareto-optimal set correspond to operating points on an optimal ROC curve, whose performances describe the limiting sensitivity-specificity tradeoffs that the classifier can provide for the given training dataset. Conventional non-evolutionary optimization techniques have not been successfully extended to the multiobjective case because they are not designed to operate on multiple solutions. Because GAs are population-based, they have formed the basis of several multiobjective optimization techniques, collectively referred to as multiobjective GAs (MOGAs) [16–19].

In this paper, we investigate the application of a MOGA called a niched Pareto GA (NP-GA) for optimizing the performance of two popular diagnostic classifiers. The paper is organized as follows. Section II contains a general introduction to automated classifiers and a brief description of the NP-GA. Section III describes the two classifiers that were studied, and it describes how the NP-GA was employed to train them. The results of the two optimizations are presented in Section IV. Sections V and VI contain a discussion of the results and a summary of the advantages and drawbacks of the proposed approach to diagnostic classifier training and ROC curve generation.

II. Background

A. Automated Classifiers

An automated binary classifier separates two classes of observations (e.g. images) and assigns new observations to one of the two classes. In this paper, we will label the two classes as normal (no disease evident) and abnormal (indicative of disease), denoted by π_n and π_a, respectively. Certain characteristics of the observations, called features, are used in making the classification decision. The set of features corresponding to an observation can be expressed by a vector $\vec{x} = [x_1, x_2, \ldots, x_p]$. In order for the classifier to be "trained," we start with a dataset of known pathology called the training dataset. A graphical depiction of an automated classifier for a two feature example ($p = 2$) is shown in Fig. 1. The $\{x_1, x_2\}$ space spanned by the feature vectors is denoted by $\mathcal{S}$. An automated classifier uses a parameter vector $\vec{w}$ to partition this space into the sets $\mathcal{C}_n(\vec{w})$, the set of observations that belong to class π_n, and $\mathcal{C}_a(\vec{w})$, the set of observations belonging to class π_a. The parameters $\vec{w}$ of a classifier can represent, for example, the weights of an artificial neural network or the threshold values in a rule-based classifier. For a fixed $\vec{w}$, $\mathcal{C}_n(\vec{w}) \cup \mathcal{C}_a(\vec{w}) = \mathcal{S}$, and $\mathcal{C}_n(\vec{w}) \cap \mathcal{C}_a(\vec{w}) = \emptyset$.

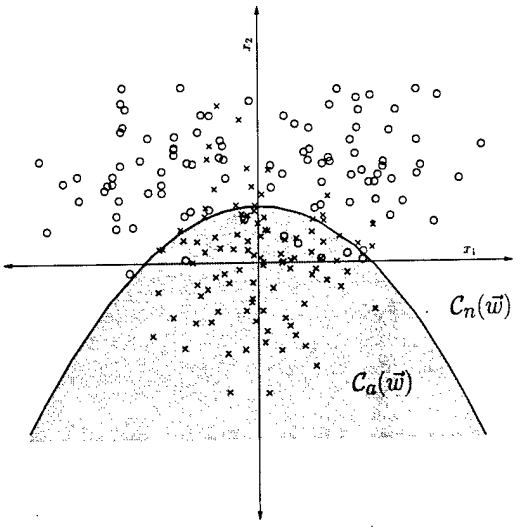


Fig. 1. The job of an automated classifier is to partition the multidimensional feature space into two partitions, $\mathcal{C}_a(\vec{w})$ belonging to class π_a and $\mathcal{C}_n(\vec{w})$ belonging to class π_n. These partitions, $\mathcal{C}_n(\vec{w})$ and $\mathcal{C}_a(\vec{w})$, are shown by the shaded and unshaded regions. The two classes, π_a and π_n, are represented by different symbols (x's and o's). The decision boundary is denoted by the solid line separating the shaded from the unshaded region.

Given a measurement $\vec{x}$, the classifier assigns $\vec{x}$ to class π_n if $\vec{x} \in \mathcal{C}_n(\vec{w})$ or to class π_a if $\vec{x} \in \mathcal{C}_a(\vec{w})$. The probability that an observation belonging to class π_a is correctly classified is referred to as the sensitivity of the classifier, denoted by $Sens(\vec{w})$. Similarly, the probability that an observation is correctly classified as belonging to class π_n is referred to as the specificity of the classifier, denoted by $Spec(\vec{w})$. Note that both the sensitivity and specificity of the classifier depend explicitly on the choice of $\vec{w}$ and implicitly on the underlying distribution of the normal and abnormal observations. The sensitivity is a measure of how well the classifier performs on abnormal cases, whereas the specificity is a measure of how well a classifier performs on normal cases. In practice, the fraction of class π_a observations that are correctly classified is used as an estimate of $Sens(\vec{w})$. Likewise, the fraction of class π_a observations that are correctly classified is used as an estimate of $Spec(\vec{w})$.

A popular construct used for describing the performance of a diagnostic classifier is the receiver operating characteristic (ROC) curve [6,7,20,21]. An ROC curve is generated by varying the value of one (or more) of the components of the parameter vector $\vec{w}$, and plotting the corresponding $Sens(\vec{w})$ and $Spec(\vec{w})$ values. For example, the output threshold is usually varied to generate an ROC curve for artificial neural networks [22]. Traditionally, the classifier is trained prior to the generation of the ROC curve [22,23]. In this situation, all but one point on the ROC curve represent operating points other than the one to which the classifier was naturally trained. An ROC curve that was generated with the same dataset that was used to train the classifier is referred to as a "consistency" ROC curve. A "validation" ROC curve is obtained when the curve is

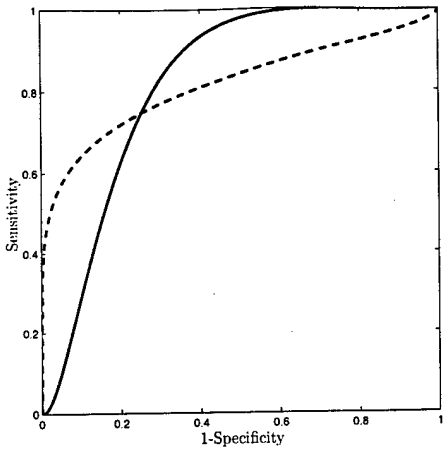


Fig. 2. The two ROC curves have equal A_z values, but, depending upon the relative preferences concerning the sensitivity or specificity of the detection task, one curve is typically preferred over the other.

generated from an independent data set, and represents an unbiased estimate of classifier performance [24]. Two typical ROC curves are shown in Fig. 2. The area under an ROC curve, or A_z, is an accepted way of comparing overall classifier performance [6, 7, 20, 21]. Two curves may have equal A_z values, as shown in Fig. 2; however, one of the curves will typically be preferred over the other, depending upon the relative preference of the sensitivity and the specificity needed for the task at hand.

For certain types of classifiers, such as rule-based systems [3,25], it may not be clear how $\vec{w}$ should be varied to sweep out the ROC curve that best represents the sensitivity-specificity tradeoffs that are achievable by the classifier on the specified dataset. The ROC curves generated by varying different sets of components of $\vec{w}$ will generally be different, representing different sensitivity-specificity tradeoffs that are possible. In this work, we demonstrate that this ambiguity can be removed if one uses the performances of the solutions returned by a multiobjective optimization of the classifier to define the ROC curve.

B. The Niched Pareto GA

We have implemented a multiobjective optimization technique called a Niched Pareto GA (NP-GA), which is described in detail by Horn *et al.* [26]. Other types of MOGAs have been proposed and are described in reference [18]. The NP-GA can be viewed as a conventional (scalar) GA that uses a modified tournament selection mechanism and ranking scheme. Readers not familiar with genetic algorithms may consult reference [9]. In the remainder of this section we review the NP-GA proposed by Horn *et al.* [26].

In order to directly address the multiobjective nature of the optimization problem, NP-GAs employ the concept of *dominance*. A solution to the optimization problem is called *non-dominated* if there is no solution superior to it in all objectives. It is the goal of the NP-GA to discover the set of all non-dominated solutions, referred to as the *Pareto-optimal set*, all of which are considered to be equally valid solutions to the problem in the absence of any *a priori* information about the relative merits of the different objectives. If a solution is not non-dominated, it is referred to as being *dominated*. A non-dominated solution is said to *dominate* a dominated solution. Equivalence classes of dominated solutions are formed by grouping them according to the number of solutions that dominate them.

This grouping of solutions into distinct classes establishes a *partial order* on the set of all solutions that is used to determine rank. We assume that the Pareto-optimal set corresponds to equivalence class 0, and that all other solutions have an equivalence class greater than zero. The rank of a particular solution is then equal to its equivalence class number. This ensures that solutions within the same equivalence class have the same rank, which reflects the fact that solutions within the same class are equally "good" in the absence of any other information.

To perform selection, the NP-GA uses a modified tournament selection method. In a scalar GA, tournament selection is one of the methods commonly used for choosing a subset of solutions in the current generation to be placed in the following generation. Implicit in its formulation is the assumption that there exists a single solution to the optimization problem; diversity among solutions in the population will be lost after a certain number of generations. This is undesirable in a multiobjective optimization where we wish to discover all of the members of the Pareto-optimal set, not simply a single solution. To circumvent this difficulty, Horn *et al.* proposed the use of a *Pareto domination tournament* in conjunction with a form of fitness sharing called *equivalence class sharing*. A Pareto domination tournament is a modified conventional tournament selection method that uses the concept of dominance to determine the winner of the tournament. First, t_{dom} randomly selected solutions are compared, and the solution with the highest rank wins (is carried over to next generation). The rank, being based on the concept of dominance, incorporates the multiobjective nature of the problem into the selection mechanism. For situations when a certain tournament size provides insufficient domination pressure, the size of the tournament (t_{dom}) can be increased.

When two or more solutions in a tournament belong to the same equivalence class (*i.e.*, have the same rank), there will not be a clear winner. A winner cannot simply be chosen at random, because genetic drift will cause the population to converge to a localized region of the Pareto-optimal set, thus obscuring other potential solutions to the optimization problem. Instead, a form of fitness sharing called equivalence class sharing is employed to determine the winner of a tied tournament. In equivalence class sharing, the winner of a tied tournament is the solution that has the smallest *niche count*. The niche count estimates the density of solutions in a localized region (niche) around a given solution. As described in reference, the niche count m_i for

the ith solution is given by

$$m_i = \sum_{j \in Pop} s(d_{ij}), \tag{1}$$

where d_{ij} is the distance (in objective space) between solutions i and j, and $s()$ is the so called *sharing function* given by $s(d) = 1 - d/\sigma_{share}$ for $d \leq \sigma_{share}$ and $s(d) = 0$ otherwise. Here, σ_{share} is called the *niche radius*, which represents the maximum distance between solutions that will result in an increase in their niche counts. By employing fitness sharing in this way, the Pareto-optimal set is more likely to be uniformly sampled, thus providing a more diverse set of potential solutions to the optimization problem from which the user can choose.

III. Methods

We trained a linear classifier and an artificial neural network by using both conventional optimization techniques and the NP-GA. Two-dimensional exclusive-OR data [27], sampled from the density functions shown in Fig. 3, were used for this study because classifiers typically have difficulty in adequately classifying both classes of data for this problem. Two-dimensional isotropic standard normal distributions with mean (μ_{x_1}, μ_{x_2}) and variance 1 were sampled in the four regions of the exclusive-OR problem. The normal class (dashed lines in Fig. 3) occupied the regions centered at $(1.3, 1.3)$ and $(-1.3, -1.3)$. The abnormal class occupied the regions centered at $(1.3, -1.3)$ and $(-1.3, 1.3)$. A total of 100 normal and 100 abnormal samples were generated for training data. An additional 10,000 normal and 10,000 abnormal samples were generated for testing the classifiers after they had been trained. The performances of the conventionally optimized and NP-GA optimized classifiers were evaluated on both the training and the testing datasets.

A. NP-GA Implementation

The NP-GA was employed to simultaneously maximize the sensitivity and specificity of a linear classifier and an ANN with a single hidden layer. The value of each component of $\vec{w}$ was restricted to remain within a maximum and minimum value, determined prior to the optimization. A binary representation of the chromosomes [9] was utilized so that each real-valued parameter in $\vec{w}$ was encoded by a binary number of fixed length. The range of each component of $\vec{w}$ and the length of its binary representation determined that parameter's floating-point precision. The encoding was accomplished by linearly scaling the floating point number using its specified range to an integer between 0 and $2^n - 1$ where n is the number of bits. Standard single-point crossover and standard mutation were employed as the genetic operations [9]. The rates of the genetic operations were determined empirically by performing multiple optimizations. A crossover rate of 30% and a mutation rate of 5% were found suitable for the problems studied. A t_{dom} value of 4 and a σ_{share} value of 0.1 (or 10% of the range of each objective) were also found to work well for the optimization problems discussed in this paper. A discussion of these parameter settings is presented later.

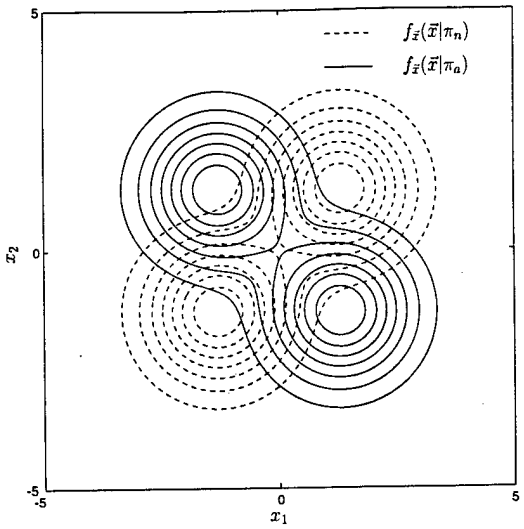


Fig. 3. Contour diagrams of the two density functions that make up the exclusive-OR problem. The abnormal class (solid lines) occupies the upper-left and lower-right quadrants, whereas the normal class (dashed lines) occupies the upper-right and lower-left quadrants.

B. Classifiers

A linear classifier attempts to separate the two classes of observations by using a linear decision boundary. We employed logistic discriminants [5] in order to implement this classification. A logistic discriminant projects the data onto a decision variable, and then a threshold is applied for determining whether a given observation belongs to π_a or π_n. The abnormal set for a logistic discriminant with parameter vector $\vec{w}$ is defined as

$$C_a(\vec{w}) = \{\vec{x} : g(\vec{x}'\vec{w}^{\mathrm{T}}) \geq 0.5\}, \tag{2}$$

where $\vec{x}' = [x_1, x_2, \ldots, x_p, -1] = [\vec{x}, -1]$ and $g()$ is a sigmoidal function with output bound between 0 and 1 [5]. The normal set is defined as $C_n(\vec{w}) = \mathcal{S} - C_a(\vec{w})$. The conventional method for generating an ROC curve for a logistic discriminant is to vary the final parameter in the vector $\vec{w}$, which results in a translation of the decision boundary.

The NP-GA was used to optimize the parameters of a logistic discriminant so as to work with the exclusive-OR data described previously. All 3 components (for 2D problems, $\vec{w}$ has 3 components) of the parameter vector $\vec{w}$ were allowed to range between -3 and 3. With a population size of 500 solutions, we ran the NP-GA for a total of 100 generations. Conventional logistic discriminant training, as described in reference [5], was employed to compare with the NP-GA results.

An artificial neural network (ANN) is a set of connected nodes that is loosely based on the human neuron system [5, 27–30]. For classification purposes, an ANN can be thought of as a mapping function that uses the weights $\vec{w}$ to map

an input vector $\vec{x}$ to a scalar quantity to which a threshold is applied to determine whether $\vec{x}$ belongs to class π_a or π_n. Unlike logistic discriminants, an ANN can separate the two classes of observations using a non-linear decision boundary. The abnormal set of observations for an ANN using the weights $\vec{w}$ is given by

$$C_a(\vec{w}) = \{\vec{x} : h(\vec{x}; \vec{w}) \geq 0.5\}, \quad (3)$$

where $h(\vec{x}; \vec{w})$ represents the non-linear mapping of the input features to the single output value bound between 0 and 1.

We applied the NP-GA to optimize an ANN on the exclusive-OR data. A two-layered ANN with 2 inputs, 2 hidden units, and one output unit was employed. This corresponded to a total of 9 parameters to be optimized. The magnitudes of the weights were forced to lie between -5 and 5 in order to simplify the optimization task and to regularize the problem somewhat, because large weight values represent complex decision boundaries [28]. A population size of 3000 solutions was run for a total of 100 generations for this study. Conventional error-backpropagation ANN training [5, 27, 29, 30] was also employed numerous times using different initial conditions. A comparison of the performances of the NP-GA results with the best conventional results will be shown along with a comparison of the NP-GA performances with a conventional optimization that was trapped in a local minimum. The conventional ROC curves were generated by varying the output bias weight value, which corresponds to one component of $\vec{w}$. This is equivalent to varying the neural network output threshold. It should be noted that Woods and Bowyer [23] studied the effect of varying weight values other than the output bias weight in generating ROC curves. Their study concluded that varying a subset of the weights can produce better ROC curves than the ROC curves produced by varying the output threshold as is conventionally done. By applying the NP-GA to ANNs, however, we are effectively allowing all the weights to vary when generating the ROC curve, including both the output threshold and the hidden layer bias weights studied in the Woods and Bowyer work.

IV. Results

A. Linear Classifier

Figure 4 shows the performances of the non-dominated solutions returned by the NP-GA and the ROC curve that resulted from the conventional training, generated by thresholding the output value. The operating points obtained by the NP-GA are seen to be better than the corresponding operating points on the conventional ROC curve in the high sensitivity region. Figure 5 demonstrates the same behavior when the NP-GA solutions and the conventional solution are evaluated on the independent data set. This is evidence that the performance improvement achieved by the NP-GA training was not simply a result of over-training. However, because the training data were sparse between the four regions of the exclusive-OR data, a few of the solutions returned by the NP-GA show slight signs of overfitting when tested on the 20,000 testing samples, as is demonstrated by the fact that a few solutions are dominated when evaluated on the test set. The majority of the solutions, however, do not show signs of overtraining.

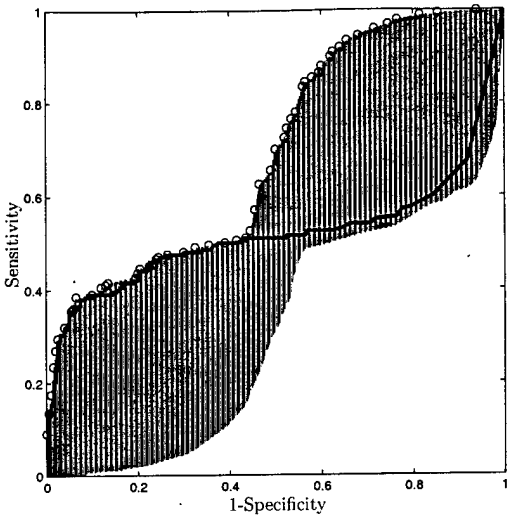


Fig. 4. Consistency results of the logistic discriminant training using exclusive-OR training data. The circles represent the performances of the non-dominated solutions returned by the NP-GA based training. The solid line is the conventional ROC curve produced by varying the output threshold value of the logistic discriminant after it was trained using a scalar optimization technique. The shaded region shows the performances achievable by all possible weight vectors $\vec{w}$.

The ROC curve for the conventionally trained logistic discriminant was generated by varying the output threshold (final parameter in $\vec{w}$) and plotting the corresponding sensitivity and specificity values. Figure 6 shows the decision boundaries at various output thresholds for the conventionally trained logistic discriminant. Decision boundaries corresponding to different threshold values are seen to be parallel. Because of this, the classifier only performs well in the low-sensitivity region. If, however, the decision boundaries were rotated by 90 degrees to those shown in Fig. 6, the classifier would, instead, perform well in the high sensitivity region. The advantage of the NP-GA is that, at different ROC operating points, the orientation of the decision boundary can be different. Thus, the NP-GA trained logistic discriminant can perform optimally in both the high and low sensitivity regions. This is because with the NP-GA, all components of $\vec{w}$ are effectively allowed to vary when generating the ROC curve rather than just varying the value of one of the parameters and keeping the other two fixed.

B. Artificial Neural Network

The performances of the NP-GA results on the 200 training samples is shown in Fig. 7. The best conventional ANN optimization ROC curve, created by varying the output threshold, is also shown in Fig. 7. The NP-GA result is either equal to or better than the best conventional result

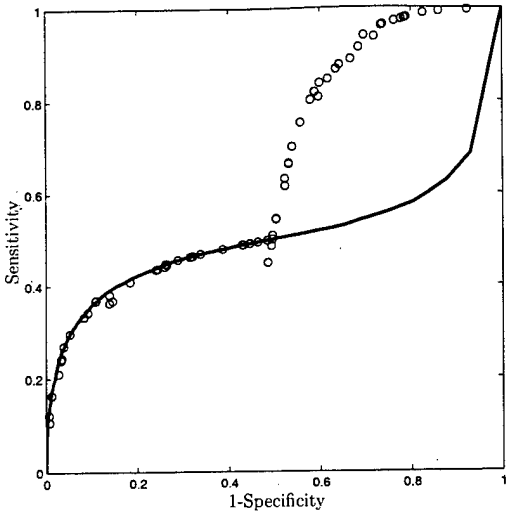


Fig. 5. Validation results of the logistic discriminant training for 20,000 samples from the exclusive-OR data distribution to evaluate the performances. The circles represent the performances of the non-dominated solutions returned by the NP-GA based training. The solid line is the conventional ROC curve produced by varying the output threshold value of the logistic discriminant after it was trained using a scalar optimization technique.

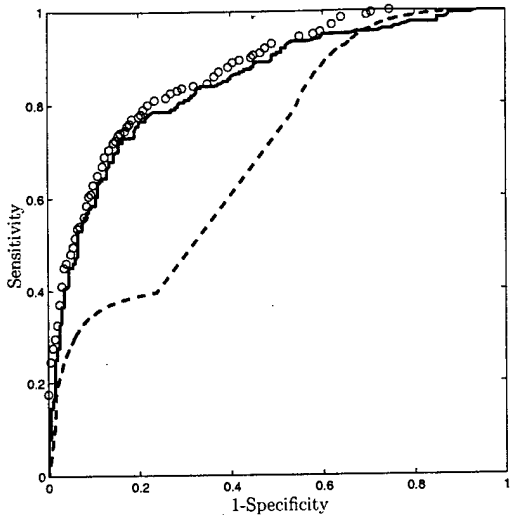


Fig. 7. Consistency results of the ANN training using exclusive-OR training data. The circles represent the performances of the non-dominated solutions returned by the NP-GA based training. The solid line is the conventional ROC curve produced by varying the output threshold value of the ANN after it was trained using a scalar optimization technique. The dashed line represents the result of a conventionally trained ANN trapped in a local minimum. The conventional training became trapped in local minima in approximately 30% of the conventional optimizations performed.

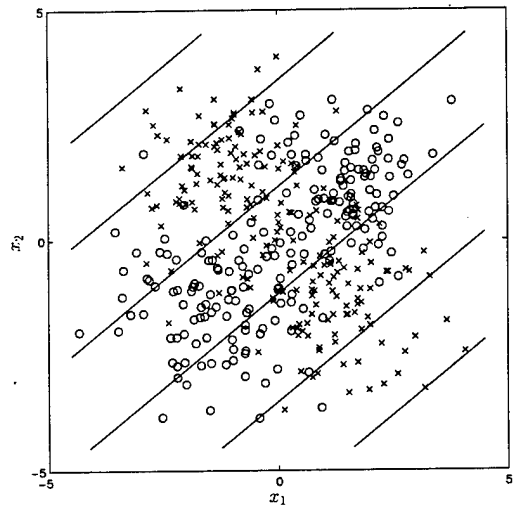


Fig. 6. An explanation of why the conventionally-trained logistic discriminant only performs well in the low sensitivity region. The decision boundaries corresponding to different output threshold values of the discriminant are shown superimposed on the data distribution. The o's represent normal signals and the x's represent the abnormal signals. The abnormal region is to the left of each decision boundary and the normal region is to the right of each decision boundary. When the threshold value is varied, the decision boundary is simply translated with its orientation remaining fixed. By analyzing the sensitivities and specificities for each decision boundary, one can generate the conventional ROC curve shown in Fig. 4. In order for the classifier to perform well in the high sensitivity region, the decision boundaries would have to be rotated by 90 degrees which would result in the classifier performing poorly in the low sensitivity region.

at all points. The differences are small in most regions, but substantial in the very high sensitivity region of the ROC curve. No regularization techniques were applied to the conventional optimization; therefore, one would typically be concerned about over-training. Figure 8 shows the validation ROC curves generated by applying the optimized results to the 20,000 testing samples. Again, the NP-GA result is closely matched with the conventional result at most places in ROC space except in the high sensitivity region where the NP-GA result is noticeably better than the conventional result. Overtraining was not a noticeable problem in both of these optimizations because the structure of the ANN was limited (2 hidden nodes) in both runs and the parameter range of the NP-GA was limited as well.

Local minima often plague conventional ANN optimizations. We found that, depending upon the initial staring point, our ANN converged to local minima about 30% of the time as was evident by comparing the ROC curves of the different ANN optimizations. The NP-GA never had a problem with local minima. Figure 7 also shows the performance of the conventional result that resided in a local minimum in the parameter space (dashed line). The NP-GA result is substantially better at almost all points in ROC space.

C. NP-GA Performance

We conducted experiments to analyze the behavior of the NP-GA and verify that our choice of NP-GA operating parameter settings was reasonable. Figure 9 demonstrates the convergence of the non-dominated set when the ANN

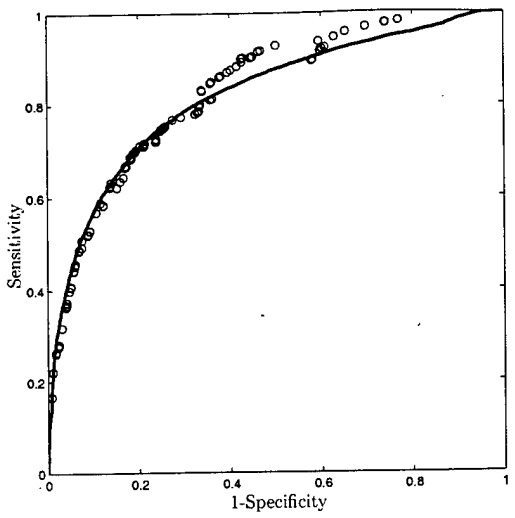


Fig. 8. Validation results of the ANN training on 20,000 samples from the exclusive-OR data distribution to evaluate the performances. The circles represent the performances of the non-dominated solutions returned by the NP-GA based training. The solid line is the conventional ROC curve produced by varying the output threshold value of the ANN after it was trained using a scalar optimization technique.

was trained using the previously described training data and operating parameter settings. Figures 9(a), 9(b), 9(c), and 9(d) show the performances of the non-dominated solutions evaluated on the training data at generations 2, 5, 13, and 100, respectively. It can be seen that the loci of operating points migrate upward and to the left as the generation number increases. Beyond 100 generations, the loci of operating points remain approximately constant, demonstrating that the NP-GA had converged to a stable set of solutions. It should also be noted that the relatively high density of the operating points returned by the NP-GA indicates that the non-dominated set of solutions was adequately sampled.

Although the data described above demonstrate that the NP-GA converged when training the ANN, we do not know whether the final set of solutions represents the best possible set of solutions (i.e., the Pareto-optimal set). To verify this, one would have to evaluate the performances of all the possible combinations of parameter values of the ANN, which is not a computationally tractable problem with current computer technology. We can, however, compute this for the linear classifier because it possesses only 3 free parameters. The shaded region in Fig. 4 shows the operating points achievable by all possible parameter settings for the linear classifier. Because most of the operating points returned by the NP-GA lie on the upper-left boundary of the shaded region, we can conclude that, for this example, the NP-GA was successful at converging to the Pareto-optimal set.

As was noticed in reference [19], we observed that the size of the Pareto dominant tournament (t_{dom}) significantly affected the convergence behavior of the NP-GA. Figure 10 shows the operating points returned by two separate applications of the NP-GA to the ANN training. The upper set of solutions, discussed previously, was obtained with $t_{dom} = 4$. The lower set of solutions was obtained using the same NP-GA operating settings except with $t_{dom} = 2$. With $t_{dom} = 2$, the NP-GA returned a set of solutions that were clearly suboptimal. One explanation of this result is the following: When a tournament selection scheme is used, there is a non-zero probability that a solution in a given population will not be selected to compete in any of the tournaments. This can result in a potentially "good" solution being lost by the NP-GA. The probability of losing a solution in this way is equal to $(\frac{N-1}{N})^{t_{dom}N}$, where N is the population size. When N is large, this probability converges to $e^{-t_{dom}}$. For $t_{dom} = 2$, this corresponds to a probability of 0.135 of losing a solution in any given population. When $t_{dom} = 4$, this probability is reduced to 0.018. By increasing the size of the tournament, we reduce the probability of losing a potentially good solution which could contribute to inadequate convergence of the NP-GA.

There are problems, however, with using too large a tournament size. When we used large values of t_{dom} (for example, $t_{dom} > 20$), the NP-GA converged to a solution similar to that achieved for $t_{dom} = 4$, but subsequently fluctuated about that solution as a function of generation number. This instability is a result of having domination tournaments in which multiple non-dominated solutions are forced to compete. When non-dominated solutions are forced to compete in multiple tournaments, one or more of the members of the non-dominated set will inevitably be lost. (The niche count determines the winner of a tied tournament.) The observed instability of the non-dominated set is a result of losing and re-gaining non-dominated solutions. When large values of t_{dom} are used, the value of the niche size (σ_{share}) becomes increasingly important, because multiple tied tournaments may arise. For $t_{dom} = 4$, we found that the NP-GA performance was relatively insensitive to the value of σ_{share}.

V. Discussion

Genetic algorithm parameters are difficult to determine, and few methods exist to systematically set the GA parameters. The total number of generations, the number of solutions in each generation, the crossover rate and the mutation rate were determined experimentally. Various GA parameter combinations were tested and the results were compared. We found a set of parameters for which the results were consistent in the sense that multiple optimizations gave solutions with similar performances. If the sets returned by different NP-GA runs were not optimal, one would expect that multiple NP-GA runs would return sets with either better or poorer performances. We also attempted to use various σ_{share} values and found that the NP-GA results were robust with respect to σ_{share}.

The NP-GA exhibits several advantages over conventional classifier training techniques. One advantage is that the objective function describing the optimization task is a vector-valued function. This eliminates completely the

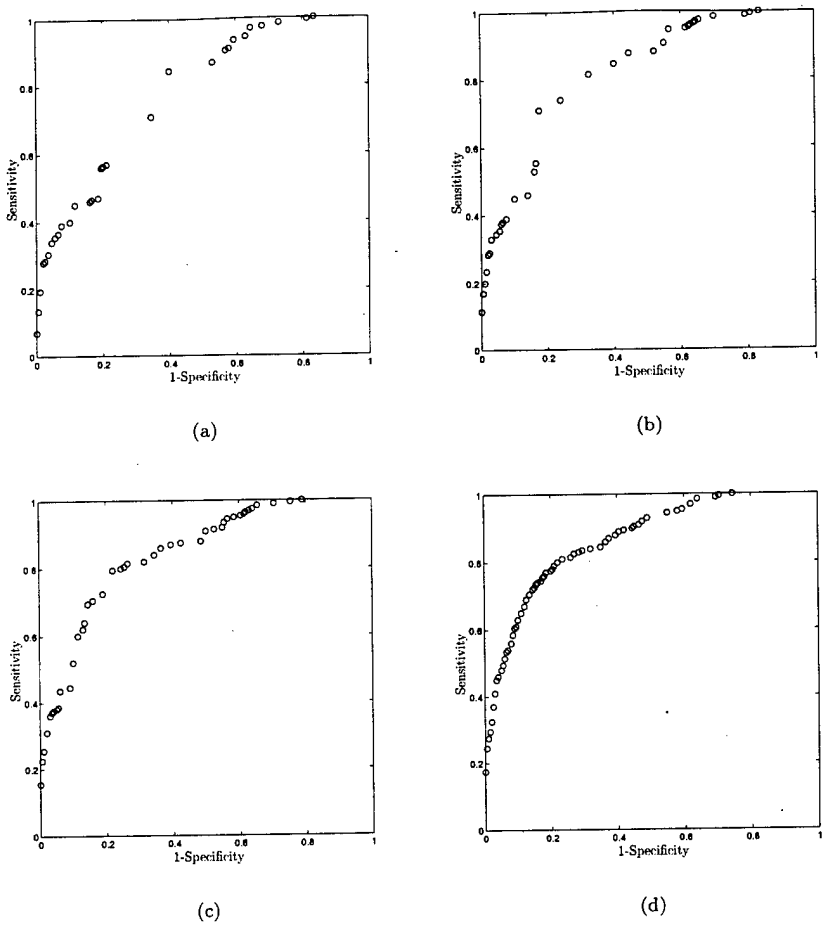


Fig. 9. Convergence of the NP-GA for the ANN training example as described in the text. Subfigures (a), (b), (c), and (d) show the performances of the non-dominated solutions at generation numbers 2, 5, 13, and 100, respectively. As the generation number increases, the loci of operating points migrate upward and to the left.

need to aggregate the different objectives (sensitivity, specificity) into a single scalar function. Rather, *a priori* information about the relative preferences of the objectives can be used post-optimization to choose a member of the Pareto-optimal set as the ultimate solution to the problem.

Another advantage is that a set of non-dominated solutions is returned, rather than a single solution. This allows one to select the solution (ROC operating point) whose performance is most clinically appropriate for the diagnostic task at hand. Conventional classifier optimizations can return a series of solutions in the form of an ROC curve obtained by varying certain components of $\vec{w}$ after the classifier has been trained. If the scalar cost function employed is an aggregation of sensitivity and specificity directly then only one point in ROC space is guaranteed to be optimal. If the scalar cost function is an aggregation of two different performance measures (such as the sum-of-squares error function for ANNs) then no point is guaranteed to be optimal in ROC space. The NP-GA circumvents this problem by allowing all parameters in $\vec{w}$ to effectively vary *in an optimal manner* when sweeping out the ROC curve. In this sense, the consistency ROC curve returned by the NP-GA, assuming that the optimization is complete, is optimal at every point. All other possible performances for the same classifier and dataset are either equal to or less than the ROC curve returned by the NP-GA optimization. Training the classifier to operate at a particular operating point and then varying a subset of the parameters in a predetermined way to generate the ROC curve does not ensure this.

As we have alluded to earlier, conventional methods of classifier optimization can, in fact, produce the Pareto-optimal operating points through multiple runs of the scalar optimization procedure with different weighting factors on sensitivity and specificity (see the Appendix for a more detailed discussion of this). Sensitivity and specificity are, however, discrete counting statistics and hence are not differentiable functions of $\vec{w}$. Conventional gradient-based optimization methods such as backpropagation cannot be employed in this situation. One is therefore left with running multiple scalar stochastic optimizations (such as

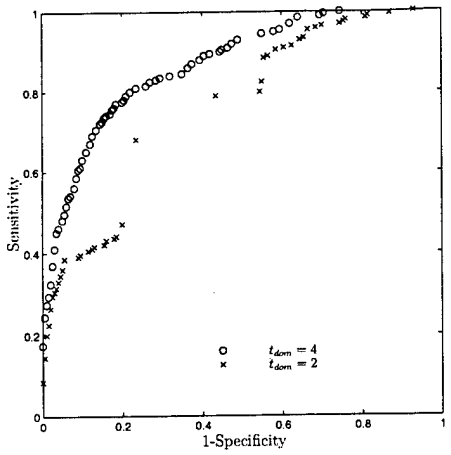


Fig. 10. Effect of t_{dom} on convergence of the NP-GA. At a t_{dom} value of two, the NP-GA converged prematurely because of the lack of domination pressure. For the problems studied in this paper, a t_{dom} value of four resulted in reliable convergence of the NP-GA. Large values of t_{dom} caused the non-dominated set to fluctuate randomly.

GAs or simulated annealing) to produce the same operating points that were produced with one single run of the NP-GA. It is also not always clear how to set the relative weightings to evenly sample the Pareto-optimal set using a scalar optimization technique. Another option would be to run multiple optimizations using a conventional cost function such as the sum-of-squares cost function with different weightings on the two objectives. No point, however, is guaranteed to be a member of the Pareto-optimal set if this type of error function is employed. By using an NP-GA to train pattern classifiers, we are directly addressing the multiobjective nature of classification problem.

If the density functions of the normal and abnormal classes ($f_n(\vec{x})$ and $f_a(\vec{x})$, respectively) are known, then the ROC curve that is produced using the likelihood ratio $LR(\vec{x}) = f_a(\vec{x})/f_n(\vec{x})$ or any monotonic transformation of the likelihood ratio as the decision variable will be the "optimal" ROC curve [20,31]. It will exhibit the best classification performances that can be achieved with the given density functions. It is often very difficult with limited datasets to estimate the density functions of the two classes of data; thus many classifiers, including those used in this paper, make no attempt to accurately estimate these distributions. The "optimal" ROC curves that have been discussed in this work are quite different. Within the limitations of the classifier employed and the dataset used for training, the ROC curves produced using the NP-GA are optimal, i.e., there is no better ROC curve that can be produced with the same training data and classifier.

There are sacrifices that are made when the NP-GA is used for classifier optimization. GAs are population-based stochastic optimization algorithms; thus, they are typically more time consuming than are deterministic algorithms. The time to optimize the linear classifier on a 400MHz Pentium II system was on the order of 3 minutes. The time to optimize the ANN on this system was about twenty minutes. In fact, for very complex systems, an NP-GA optimization may be impractical with current computer technology. For ANNs with a large number of inputs and hidden nodes, the NP-GA may not be suitable for training with current computer technology because of the large number of parameters. In these situations, techniques for sweeping out ANN ROC curves proposed by Woods and Bowyer [23] may be better suited. The NP-GA, however, can readily be made to run in parallel which would substantially decrease the execution time.

This paper has dealt with binary classifiers. It is often important, however, to classify observations into more than two classes (benign, malignant, and normal, for example). For a three-class system, aggregating the multiple objective functions into a single scalar function suffers from the same problems as the two-class problem, but to a greater degree. Here, it is even more difficult to adequately incorporate the class preferences in the aggregated objective function. The ability of the NP-GA to circumvent this difficulty is very attractive. Because the non-dominated set of solutions will be larger, care must be taken in determining the NP-GA parameter settings to ensure that the Pareto-optimal set is adequately sampled.

Complexity and over-training are issues of great importance in diagnostic classifier research, and in particular in ANN training [28,32]. In practice, there is typically a limited amount of training data available, and some sort of regularization is imposed during the classifier training to ensure that it performs well on other (unknown) data sets. It is well known that large ANN weights correspond to complex separation functions [28,32] that may be indicative of over-training. To avoid this, we have imposed limitations on the magnitudes of the ANN weights when using the NP-GA to determine the weight values. More systematic methods of regularizing the NP-GA based training may be possible, however. One such method is to add a third component to the vector objective function that measures complexity. In this way, one can maximize the sensitivity and specificity while minimizing the complexity of the classifier. Depending on the amount and quality of the available training data, a non-dominated solution returned by the NP-GA can be chosen such that the classifier performance and generalizability of the result are appropriate for the classification task. We are currently investigating this approach to classifier training.

VI. Conclusions

We have studied the use of a niched Pareto genetic algorithm in training two popular diagnostic classifiers. Unlike conventional classifier training techniques that formulate the problem as the solution to a scalar optimization, the NP-GA explicitly addresses the multiobjective nature of the training task. It has been demonstrated that the multiobjective approach removes the ambiguity associated with defining a scalar measure of classifier performance, and that it returns a set of optimal solutions that are equivalent in the absence of any information regarding the preference of

the objectives (sensitivity, specificity). The performances of these solutions can be interpreted as operating points on an optimal ROC curve, describing the limiting trade-offs between sensitivity and specificity that are achievable by that classifier, given the available training data. The task of classifier optimization and ROC curve generation are combined into a single task. It was demonstrated that constructing the ROC curve in this way may result in a better ROC curve than is produced by conventional methods of ROC curve generation. The NP-GA optimization typically requires more computation time than do conventional non-stochastic optimization methods, which may limit its application to certain problems. The advantages of the NP-GA approach to classifier training become more pronounced when the number of classes to be classified increases beyond two.

APPENDIX

In this work, we have investigated the use of a multiobjective optimization algorithm to train diagnostic classifiers and generate ROC curves. In fact, scalar optimization methods can theoretically arrive at the same ROC curves as a multiobjective optimization. Consider the following scalar optimization problem:

$$\text{Maximize} \quad \sum_{i=1}^{p} \lambda_i f_i(\vec{w}), \tag{A1}$$

where $\vec{w}$ is an element of the space of possible parameter vectors $\mathcal{W}$, $\lambda_i > 0$ are fixed, and $\sum_{i=1}^{p} \lambda_i = 1$. Geoffrion [33] proved the following lemma:

Lemma 1: (a) If $\vec{w}_0$ maximizes Eqn. A1, then $\vec{w}_0$ is also Pareto-optimal in the vector objective space $[f_1(\vec{w}), f_2(\vec{w}), \ldots, f_p(\vec{w})]$.

(b) Let $\mathcal{W}$ be a convex set, and let the f_i be convex on $\mathcal{W}$. Then $\vec{w}_0$ is Pareto-optimal if and only if $\vec{w}_0$ maximizes Eqn. A1 for some $\lambda_i > 0$ and $\sum_{i=1}^{p} \lambda_i = 1$.

Because the multiobjective training problem as we have formulated it satisfies the convexity conditions used in the Lemma, it must be true that the optimal ROC operating points can be obtained by performing multiple scalar optimizations with varying λ_i's.

It is clear from Fig. 4 that the solutions returned by the NP-GA are Pareto-optimal because for this problem, we can plot the performances of all possible solutions (the shaded region in Fig. 4). However, in Fig. 7, we cannot plot the performances of all possible solutions due to the large dimensionality of the parameter space. We can, however, make a comparison between the solutions returned by the NP-GA and the solutions returned by multiple scalar optimizations which maximize

$$\lambda Sens(\vec{w}) + (1 - \lambda) Spec(\vec{w}) \tag{A2}$$

with λ varying between 0 and 1. We implemented a scalar GA using the same GA parameters and parameter restrictions as imposed on the NP-GA to optimize Eqn. A2. As described above, the solutions to both of these problems should be Pareto-optimal in ROC space assuming the optimizations are complete. Figure A1 compares the NP-GA solutions and the solution achieved through multiple runs of a scalar optimization with varying λ. The points returned by the multiple scalar optimizations are similar to certain points returned by the NP-GA. Note that the multiple scalar optimized solutions are clumped together in certain areas of the ROC space. It is unknown, *a-priori*, how to vary λ to evenly sample the Pareto-front, whereas the NP-GA employs niching to ensure an even sampling of the Pareto-front or optimal ROC curve. One also cannot employ gradient-based techniques to optimize discrete performance measures such as sensitivity and specificity. Because of this, it was necessary to perform 20 separate stochastic scalar optimizations to get the 20 ROC operating points. On the other hand, a more complete sampling of the ROC curve was obtained by a single run of the NP-GA, which required approximately the same CPU time as one run of the scalar optimizer. So despite the theoretical equivalence of the two methods, there are practical advantages to performing a single multiobjective optimization over multiple scalar optimizations.

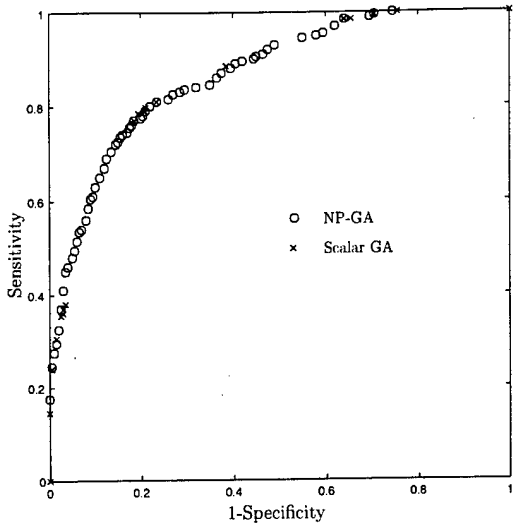


Fig. A1. A comparison of the solutions returned by the NP-GA and the solutions returned by 20 scalar optimizations employing a weighted sum of sensitivity and specificity as the scalar cost function. The two methods returned many similar solutions, but the solutions returned from multiple scalar optimizations tended to clump together in certain areas whereas the NP-GA solutions were uniformly distributed in ROC space. Note that only 18 of the 20 scalar solutions were distinct.

ACKNOWLEDGMENTS

The authors thank Dr. Charles E. Metz and Darrin Edwards for their many helpful suggestions. The authors also thank Dr. Xiaochuan Pan and Dr. Maryellen L. Giger for their frequent encouragement. This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-

1-7202) and USPHS grants CA24806 and RR11459.

REFERENCES

[1] M. L. Giger, "Computer-aided diagnosis," *RSNA Categorical Course in Physics*, pp. 283–298, 1993.

[2] K. Doi, M. L. Giger, R. M. Nishikawa, K. R. Hoffmann, H. MacMahon, R. A. Schmidt, and K.-G. Chua, "Digital radiography: A useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images," *Acta Radiologica*, vol. 34, pp. 426–439, 1993.

[3] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer-aided detection of clustered microcalcifications on digital mammograms," *Medical and Biological Engineering and Computing*, vol. 33, pp. 174–178, 1995.

[4] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics, New York: Springer-Verlag Inc., 1996.

[5] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press, 1995.

[6] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, no. 4, pp. 283–298, 1978.

[7] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720–733, 1986.

[8] M. A. Anastasio, H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi, "A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms," *Medical Physics*, vol. 25, no. 9, p. 1613, 1998.

[9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1989.

[10] D. B. Fogel, E. C. Wasson III, and E. M. Boughton, "Evolving neural networks for detecting breast cancer," *Cancer Letters*, vol. 96, pp. 49–53, 1995.

[11] J. D. Shaffer, D. Whitley, and L. J. Eshelman, "Combinations of genetic algorithms and neural networks: A survey of the state of the art," in *COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks* (L. D. Whitley and J. D. Shaffer, eds.), (Los Alamitos, California), IEEE Neural Networks Council, IEEE Computer Society Press, 1992.

[12] B. Sahiner. H.-P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler. and M. M. Gootsitt, "Image feature selection by a genetic algorithm: Application to classification of mass an normal breast tissue." *Medical Physics*, vol. 23, no. 10, p. 1671, 1996.

[13] Y. Yuan and H. Zhuang. "A genetic algorithm for generating fuzzy classification rules," *Fuzzy Sets and Systems*, vol. 84, no. 1, 1996.

[14] R. Srikanth. R. George, N. Warsi, D. Parbhu, F. E. Petry, and B. P. Buckles, "A variable-length genetic algorithm for clustering and classification," *Pattern Recognition Letters*, vol. 16, no. 8, p. 789, 1995.

[15] D. White and P. Ligomenides, "GANNet: A genetic algorithm for optimizing topology and weights in neural network design," *Lecture Notes in Computer Science*, no. 686, pp. 322–327, 1993.

[16] C. M. Fonseca and P. J. Fleming, "Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization." in *Genetic Algorithms: Proceedings of the Fifth International Conference* (S. Forrest, ed.), (San Mateo, CA), Margan Kaufmann, July 1993.

[17] J. D. Schaffer and J. J. Grefenstette, "Multi-objective learning via genetic algorithms," in *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 593–595, Margan Kaufmann, 1985.

[18] C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization," *Evolutionary Computation*, vol. 3, no. 1, pp. 1–16, 1995.

[19] J. Horn and N. Nafpliotis, "Multiobjective optimization using the niched pareto genetic algorithm," in *Proceeding of the First IEEE Conference on Evolutionary Computation*, vol. 1, (Piscataway, NJ), pp. 82–87, IEEE World Congress on Computational Intelligence, IEEE Service Center, 1994.

[20] J. Egan, *Signal Detection Theory and ROC Analysis*. Academic Press, New York, 1975.

[21] J. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, 1979.

[22] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, no. 1, pp. 81–87, 1993.

[23] K. Woods and K. W. Bowyer, "Generating ROC curves for artificial neural networks," *IEEE Transactions on Medical Imaging*, vol. 16, no. 3, pp. 329–337, 1997.

[24] C. E. Metz, "Evaluation of digital mammography by ROC analysis," in *Digital Mammography* (K. Doi, ed.), International Congress Series, pp. 61–68, Elsevier, 1996.

[25] M. A. Anastasio, M. A. Kupinski, and R. M. Nishikawa, "Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 1089–1093, 1998.

[26] J. Horn and N. Nafpliotis, "Multiobjective optimization using the niched parto genetic algorithm," illiGAL report no. 93005, University of Illinois at Urbana-Chanpaign, July 1993.

[27] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks*. Reading, Massachusetts: Addison-Wesley, 1989.

[28] D. J. S. MacKay, *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, Pasadena, California, 1992.

[29] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley Publishing Company, 1991.

[30] S. Haykin, *Neural Networks, A Comprehensive Foundation*. New York, NY: Macmillan, 1994.

[31] H. L. Van Trees, *Detection, estimation, and modulation theory (Part I)*. New York: Academic Press, 1968.

[32] M. A. Kupinski and M. L. Giger, "Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms," in *Proceedings of the 19th International Conference of Engineering in Medicine and Biology*, (Chicago, IL), pp. 1336–1339, IEEE/EMBS, Oct. 30-Nov. 2 1997.

[33] A. M. Geoffrion, "Proper efficiency and the theory of vector maximization," *Journal of Mathematical Analysis and Applications*, vol. 22, pp. 618–630, 1968.

Feature Selection and Classifiers for the Computerized Detection of Mass Lesions in Digital Mammography*

Matthew A. Kupinski
Maryellen L. Giger
Kurt Rossmann Laboratories for Radiologic Image Research
Department of Radiology, MC2026
The University of Chicago
5841 South Maryland Avenue
Chicago, IL 60637
m-kupinski@uchicago.edu
m-giger@uchicago.edu

Abstract

We have investigated various methods of feature selection for two different data classifiers used in the computerized detection of mass lesions in digital mammograms. Numerous features were extracted from abnormal and normal breast regions from a database consisting of 210 individual mammograms. A stepwise method, a genetic algorithm and individual feature analysis were employed to select a subset of features to be used with linear discriminants. Similar techniques were also employed for an artificial neural network classifier. In both tests the genetic algorithm was able to either outperform or equal the performance of other methods.

1. Introduction

Computer-aided diagnosis in digital mammography is a topic that has received much attention recently[4, 10, 13, 16] due to the potential benefits of double reading in mammography.[9, 12] Many computerized mass detection schemes employ a classifier, such as a neural network, to distinguish between lesions and false-positives.[5, 6, 14] At the University of Chicago, a computerized scheme is being developed in which features are extracted from potential lesion sites and merged into a single decision variable using a classifier. Numerous features can be extracted from potential lesions sites[5] making it difficult to optimally choose representative features to be used as inputs to a classifier. In this paper we will undertake the problem of feature selection for two different classifiers using a data set consisting of features extracted from lesions and false-positive detections.

2. Materials and Methods

The database employed in this study consists of 210 (53 cases) individual mammograms with 111 visible lesions. All but one of the 53 cases contains the four standard mammographic views. Images were scanned on a Konica film digitizer to a matrix size of 512 by 512 pixels with 10-bit quantization. All lesions in this database were biopsy confirmed. 302 false-positive regions were selected from the database to be used in this study along with the true-positive regions. A total of 71 features are extracted from each lesion and false-positive. A full discussion of the methods we utilized to locate and extract the features from the regions of interest (ROIs) can be found in our previously published papers.[1, 5, 15]

Linear discriminants, namely Fisher discriminants, were employed as the initial classifier. Three methods of feature selection were tested for linear discriminants. The first was to select those features that exhibited the greatest individual separation. This method of feature selection is a rough first-approximation of an optimal subset of features to be input to a linear classifier. Inter-feature correlations are not taken into account. Also, it does not provide any means of selecting the

*To appear in the 1997 proceedings of the IEEE International Congress on Neural Networks.

number of features to be used as inputs to the linear discriminant.

A second method employed to select a subset of features for a linear classifier was the stepwise selection method.[3] The performance of a set of features is characterized by a parameter known as the Wilks' lambda which is the proportion of the total variance attributed to within-group variations in the final decision variable.[3] If the Wilks' lambda is 0 then all of the variance is due to between-groups variations so the means of the two classes are well separated. Conversely, if the Wilks' lambda is 1 then all of the variance is due to within-group variations and one can conclude that the means of the two classes in the final decision variable are equal. In the stepwise method, the first feature is selected based on the Wilks' lambda of each individual feature. Successive features are selected based on the improvement of the Wilks' lambda. After a feature is added, all features are tested for removal. This continues until the statistical significance of adding or removing a feature is small. The advantages of this method is that it implicitly takes the correlation of features into account and also selects the number of features to be used as inputs.

The third method employed to select an optimal subset of features for a linear classifier was a genetic algorithm.[2, 11] A genetic algorithm (GA) is a stochastic-based search method based on the principles of evolution in nature. The fitness function we employed was the Wilks' lambda. Runs of 700 generations were made with a 0.5% probability of mutation and a 70% probability of crossover. Figure 1 shows the typical performance of the genetic algorithm. The set of features with the best performance at the end of the GA run was used as the selected input feature set. All feature sets resulting from the different selection methods were evaluated using ROC analysis and the performance was characterized by the area, A_z, under the ROC curve.[7, 8]

A three-layered back-propagation neural network was also studied as a classifier. ROC analysis was performed on both the consistency and cross validation results. Features were selected as inputs to the artificial neural network using three methods similar to those used for selecting features for the linear classifier. The features that exhibited the greatest individual separation were used in a neural network. In the second method, a forward selection method was used. The utility function was the A_z from a consistency test of a simplified ANN structure. Forward selection begins with the one feature which has the best individual performance and tests all possible combinations of that feature with another. This process continues until the

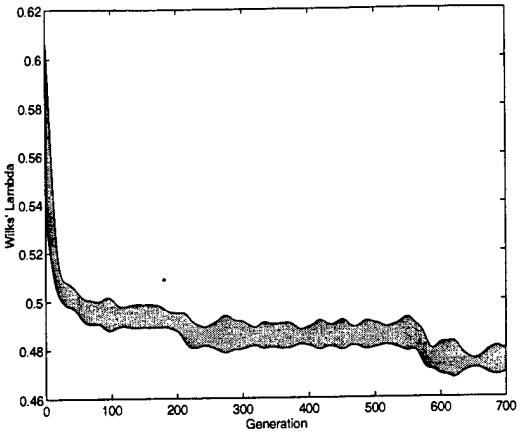


Figure 1. A typical genetic algorithm run. The shaded area represents the variation in population performance (mean ± one standard deviation) at each generation.

number of features desired has been selected. There is no feature elimination stage in the forward selection method as there is in the stepwise method. The final results from the forward selection method are input to an ANN with 2 hidden units and both consistency and cross validation results are evaluated. A genetic algorithm was also studied with the A_z from the consistency test of a simplified ANN structure used as the fitness value. A simplified structure should reduce the effects of over-fitting which often plague artificial neural networks. This final set of features resulting from the genetic algorithm were then input to a more complex ANN structure (2 hidden units) where both consistency and cross validation tests were employed.

3. Results

Table 1 shows the A_z values for the feature selection methods used for determining the inputs for a linear discriminant. Wilks' lambdas are also shown. It is clear from the table that selecting features based on their individual performance is inadequate. In Figure 2 the three different feature selection methods are compared using the ROC curves when 9 features are selected by each method. The A_z values for the feature sets selected by the genetic algorithm and the stepwise method are statistically significantly ($p < 0.05$) better than that of the single feature analysis method. The genetic algorithm shows a slight advantage over the stepwise selection method but it is not statistically

Table 1. Summary of results from the feature selection methods for linear discriminants.

Method	A_z	Wilks' Lambda	Number of Features
Single Feature Analysis	0.93	0.53	9
	0.92	0.53	10
	0.93	0.51	11
	0.94	0.50	12
Stepwise	0.94	0.47	9
Genetic Algorithm	0.95	0.47	9
	0.95	0.47	10
	0.95	0.46	11
	0.95	0.46	12

Table 2. Summary of results from the feature selection methods for artificial neural networks.

Method	Cross Validation A_z	Number of Features
Single Feature Analysis	0.96	11
Forward Selection	0.97	11
Genetic Algorithm	0.98	10

significant ($p = 0.23$).

Table 2 shows preliminary results from the ANN feature selection methods. It should be noted that multiple genetic algorithm runs were required meaning that the genetic algorithm did have trouble with local maxima. This might suggest that the probability of mutation be increased, as well as the population size, to allow for more diversity throughout the runs. As the table shows the set of features selected by the genetic algorithm was able to outperform the other two methods but the results were not statistically significant ($p = 0.06$ for the individual analysis selector and $p = 0.15$ for the forward selector). The corresponding ROC curves are shown in Figure 3.

4. Discussion

The purpose of this paper has been to introduce feature selection methods and compare their utility with two different classifiers. The results from the linear discriminant analysis show that the genetic algorithm feature selection method is as good if not better than the stepwise method. Similar results were obtained for the artificial neural network classifiers but the results

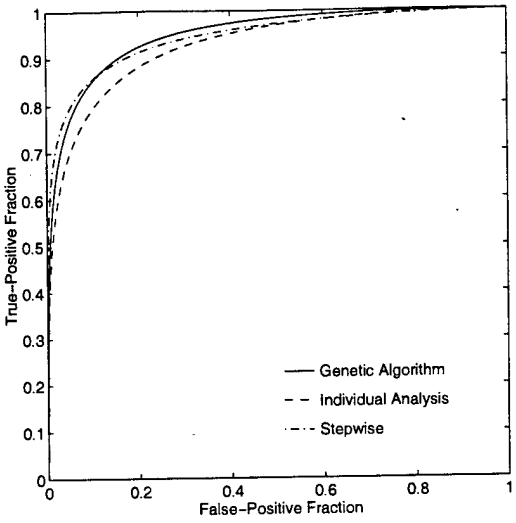


Figure 2. ROC curves for the three different linear discriminant features selection methods when 9 features were selected by each.

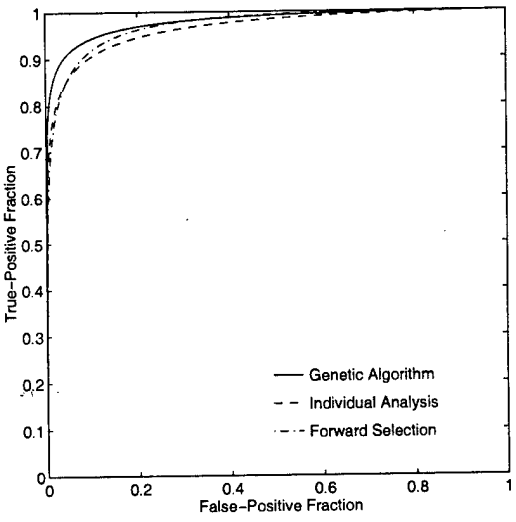


Figure 3. Cross validation ROC curves for ANN feature selectors.

were not as strong. As with all studies employing neural networks, it is possible that there is over-fitting of the data. We attempted to minimize this effect by simplifying the structure of our networks and by employing cross validation or leave-one-out tests. Future work will include investigations performed on larger data sets.

5. Acknowledgments

This work was funded in part by the US Army Medical Research and Material Command (DAMD 17-96-6058) and USPHS grant number RR11459. The contents of this exhibit are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations.

M. L. Giger is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly, actual or potential significant financial interests which would reasonably appear to be directly and significantly affected by the research activities.

References

[1] M. L. Giger. Computer-aided diagnosis. *RSNA Categorical Course in Physics*, pages 283–298, 1993.

[2] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Publishing Company, Inc., Reading, 1989.

[3] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall Inc., New Jersey, 1992.

[4] N. Karssemeijer and G. M. te Brake. Detection of stellate distortions in mammograms. *IEEE Transactions on Medical Imaging*, 15(5):611–619, 1996.

[5] M. Kupinski, M. L. Giger, P. Lu, and Z. Huo. Computerized detection of mammographic lesions: Performance of artificial neural netowkr with enhanced feature extraction. In *SPIE Medical Imaging*, volume 2434, pages 598–605, 1995.

[6] J. Y. Lo, J. A. Baker, P. J. Kornguth, and J. C. E. Floyd. Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features. *Academic Radiology*, 2:841–850, 1995.

[7] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4):283–298, 1978.

[8] C. E. Metz. ROC methodology in radiologic imaging. *Investigative Radiology*, 21, 1986.

[9] C. E. Metz and J.-H. Shen. Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. *Medical Decision Making*, 12:60–75, 1992.

[10] N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. *IEEE Transactions on Medical Imaging*, 15(1):59–67, 1996.

[11] P. Sutton and S. Boyden. Genetic algorithms: A general search procedure. *American Journal of Physics*, 62(6):549–552, 1994.

[12] E. L. Thurfjell, K. A. Lernevall, and A. A. Taube. Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191:241–244, 1994.

[13] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt. Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis. *Medical Physics*, 22(9):1501–1513, 1995.

[14] Y. Wu, M. L. Giger, K. Doi, C. E. Metz, R. M. Nishikawa, C. J. Vyborny, and R. A. Schmidt. Application of artificial neural networks in mammography for the diagnosis of breast cancer. In *SPIE Medical Imaging*, volume 1778, pages 19–27, 1992.

[15] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt. Computerized detection of masses in digital mammograms: Analysis of bilteral subtraction images. *Medical Physics*, 18(5):955–963, 1991.

[16] M. Zhang and M. Giger. Automated detection of spiculated lesions and architectural distortions in digitezed mammograms. In *SPIE Medical Imaging*, volume 2434, pages 846–854, 1995.

Classification of Suspect Regions in the Computerized Detection of Mass Lesions in Mammography

Matthew A. Kupinski, and Maryellen L. Giger

Department of Radiology, MC2026

The University of Chicago

5841 South Maryland Avenue

Chicago, IL 60637

This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202).

September 23, 2000

Abstract

In this work, we study and evaluate the ability of two classification methods (Bayesian artificial neural networks (Bayesian ANNs) and rule-based classifiers trained using a multiobjective approach) to distinguish between malignant masses and false candidates in the computerized detection of mass lesions in mammography. A Bayesian ANN is known to accurately approximate the ideal observer but may require few input features due to training dataset limitations. The multiobjective approach to classifier training returns the best possible ROC curve that a given classifier can achieve on a given training dataset. Because of overtraining issues, the multiobjective approach is typically used to optimize simple classifiers (*i.e.*, classifiers with relatively few parameters) and, thus, one can employ more input features with these classifiers. Both Bayesian ANNs and a simple rule-based classifier trained using the multiobjective approach performed well in the task of distinguishing between mass lesions and false candidates in mammography. A Bayesian ANN with 5 features outperformed the rule-based classifier which used a total of eight features. However, the rule-based classifier employed only 8 classifier parameters instead of the Bayesian ANN's 71 parameters (or ANN weights).

Keywords

Mass detection, Bayesian neural networks, multiobjective training approach, computer-aided diagnosis

I. INTRODUCTION

Breast cancer is the most common malignancy in women and the second most common cause of death from malignancy in this patient population. In the United States, more than 180,000 women develop the disease each year [1] and women who live to an advanced age have a greater than one in nine chance of developing breast cancer during their lifetimes [2]. The disease, therefore, represents a major public health problem.

Mammography, x-ray imaging of the breast, is currently the best method for the early detection of breast cancer. Between 10 and 30% of women who have breast cancer and undergo mammography have negative mammograms, however [3–8]. In approximately two-thirds of these false-negative mammograms, the radiologist failed to detect a cancer that was evident retrospectively [6,7,9,10]. The missed detections may be due to the subtle nature of the radiographic findings (*i.e.*, low conspicuity of the lesion), poor image quality, eye fatigue, or oversight by the radiologists. It has been suggested that double reading (by two radiologists) may increase sensitivity [11–15]. Thus, one aim of CAD is to increase the efficiency and effectiveness of screening procedures by using a computer system, as a "second reader" (like a "spell checker"), to indicate locations of suspicious abnormalities in mammograms as an aid to the radiologist leaving the final decision regarding the likelihood of the presence of a cancer and patient management to the radiologist [16]. The interpretation of screening mammograms lends itself to CAD since it is a repetitive task involving mostly normal images.

Pattern classification has an important role in computerized detection/diagnosis schemes in medical imaging. Thus, the ability to accurately and robustly classify suspicious image regions is vital. We previously presented a method of classification that can accurately approximate the ideal observer given a large enough training dataset [17], investigated the effect of limited datasets on feature selection [18], and also introduced a method of optimally designing and evaluating simple classifiers to be used when the training dataset is not large [19]. In this paper, we will use the knowledge gained from the research performed on feature selection, Bayesian ANNs, and the multiobjective approach to classifier training to design classifiers for a computerized mass detection method. Two types of classifiers were designed (a rule-based classifier and an ANN classifier) and evaluated using

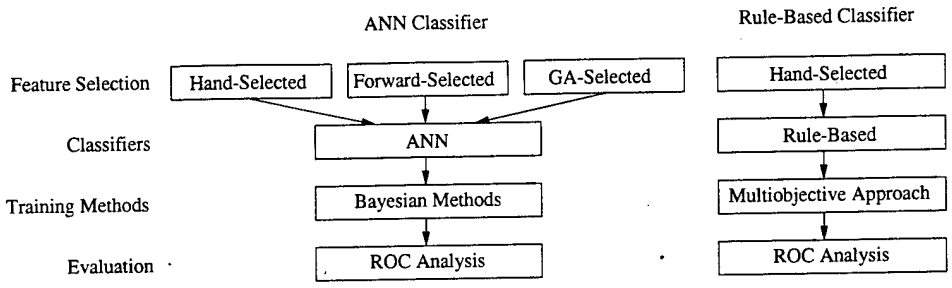
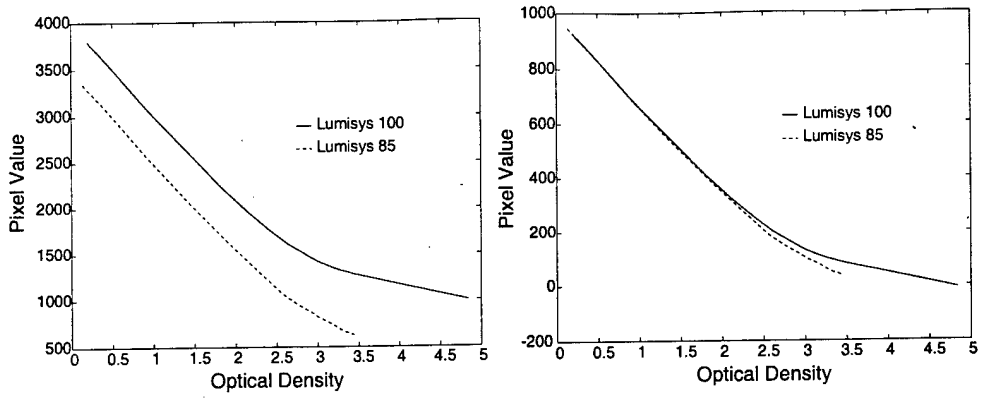


Fig. 1. An overview of the classification and feature selection methods evaluated for use in a computerized mass detection method.

various subsets of features. Figure 1 shows the various feature selection methods, classifiers, training methods, and methods of evaluation used here. This paper begins with an introduction to the mass detection method currently being developed at The University of Chicago. Section II discusses the databases used to determine the classifier parameters and to validate the classifiers. Section IV describes the features selected for the Bayesian ANN and the application of the Bayesian ANN to the mass detection method (the left side of Fig. 1). Section V describes the classifier implemented and the features selected for use in the multiobjective approach to classifier training for distinguishing between malignant lesions and false-positive candidates (the right side of Fig. 1). Finally, Sections VI and VII summarize the results and gives the overall performance of the mass detection method.

II. Database

A database of 177 screening mammography cases containing at least one malignant mass lesion and 75 cases not containing a mass were employed in this study for a total of 252 cases. There were a total of 181 malignant mass lesions which corresponded to 333 radiographically visible mass lesions on the 864 digitized films (most, but not all, of the 252 cases had 4 films available per case). The images were digitized on either a Lumisys 100 or a Lumisys 85 digitizer to 100 μm pixel size and 12-bit gray-level quantization. Different screen-film systems as well as different exposure conditions were used for the images in this database. The only image correction made was to ensure that the relationship between pixel values and optical density was approximately the same for each image. Figure 2 shows the characteristic curves of the two digitizers both before and after the correction.



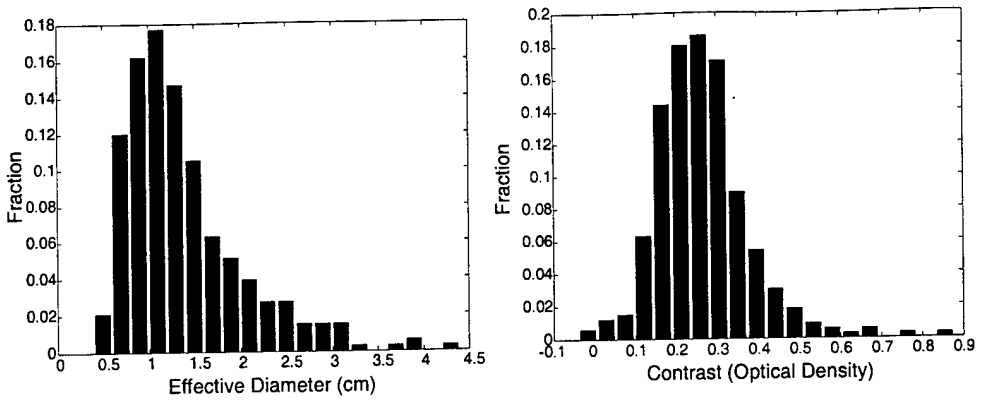
(a) (b)

Fig. 2. The original characteristic curves of the two digitizers (Lumisys 100 and Lumisys 85) are shown in (a). Images digitized on the two Lumisys digitizers were scaled so that the effective characteristic curves are similar as is shown in (b). Note that the scale on the two graphs are different; the scaled images have effectively 10 bits per pixel.

As we will discuss in the next section, however, many of the features employed were either geometric based and, hence, did not depend on the pixel values, or were gradient based features which only depend on the ratio of pixel values and not the raw values themselves.

Figure 3 shows the distribution of sizes and contrast values of the visible mass lesions as measured by a radiologist's outlined truth. The contrast was measured by taking the average within-lesion pixel value and subtracting it from the average pixel value in a region outside the lesion [20]. For these contrast calculations, the region outside the lesion was defined by pixels not inside the lesion but within a bounding box around the lesion that was extended 3 mm (or 10 pixels) in all four directions making this analogous to the "small window" case in reference [20]. Many of the lesions in the database are less than 1.5 cm in effective diameter and have a contrast less than 0.25. Thus, a substantial fraction of the lesions in this database are either small, low contrast, or both.

A total of 235 true-positive candidates were both detected by the RGI filtering method (to be discussed later) and segmented with an overlap fraction greater than 0.2 when compared to the radiologist's outlined truth. This corresponds to 150 mass cases being detected out of the 177 total mass cases. There also were over 11,000 false candidates



(a) (b)

Fig. 3. The distribution of visible mass lesion (a) sizes and (b) contrast values as calculated using the radiologist's outlines of each visible lesion.

returned by the initial detection algorithm. To ensure an independent evaluation of the mass detection method, the database was randomly split by-case into a training dataset and a validation dataset. The training dataset consisted of 126 cases with 121 visible mass lesions (or 76 mass cases) that were both detected by the initial detection algorithm and segmented with an overlap greater than 0.2. The RGI filtering algorithm also returned 5472 false candidates for the images in the training dataset from which a subset of 242 false candidates was randomly selected for classifier training. The validation dataset consisted of the remaining 126 cases with 114 visible mass lesions (or 74 mass cases) that were both detected by the initial detection algorithm and segmented with an overlap greater than 0.2 as well as 5685 false candidates.

III. Lesion Detection and Feature Extraction

Numerous research groups have developed computerized mass detection methods which use various classification techniques such as neural networks [21–24], linear discriminant analysis [23, 25], and classification trees [26] as well as various features such as spiculation [27–29], shape [30, 31], and texture [32, 33]. Rationale and details of these techniques and others can be found in various papers and chapters [16, 34–38] and in proceedings of the International Workshops on Digital Mammography [39–41] or the International Workshop

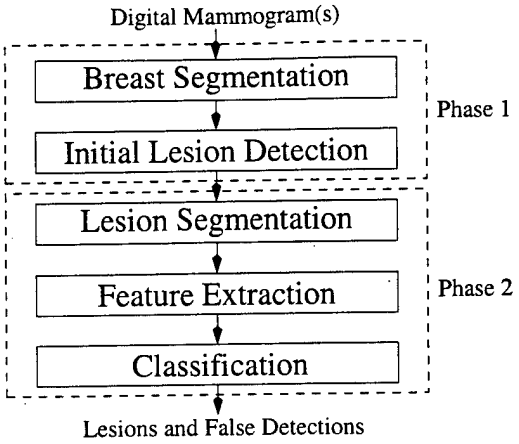


Fig. 4. Overview of a computerized mass detection scheme.

on Computer-Aided Diagnosis [42]. Many computerized mass detection methods use a similar two-phase detection approach. First, candidate lesions are located and then each candidate is further analyzed in a feature analysis and classification phase to determine the final classification of each candidate. The mass detection method being developed at The University of Chicago follows this pattern as is shown in Fig. 4. Each stage in Fig. 4 will now be briefly discussed.

A. Breast Segmentation

In order to limit the region of search for lesion detection, the breast region is initially segmented from the image. Computer-defined unexposed and direct-exposure image regions are used to generate a border around the breast region [31]. Once the breast region is known, preprocessing and search are limited by the breast border.

B. Initial Detection

The computerized mass detection method being developed at The University of Chicago is a two-phase detection method; the first detection phase locates suspicious areas called lesion candidates. The methods used in this phase are typically highly sensitive but return an unacceptable number false-positive candidates to be used alone. We have developed a non-linear filtering technique called radial-gradient index (RGI) filtering [43] to be used as the initial detection algorithm. RGI filtering takes as input the original mammogram and

the breast segmentation results returned by the breast segmentation stage and outputs a filtered image representing the confidence that a lesion is present at each location within the image.

The overall performance of the RGI filtering technique is shown in Fig. 5 with the implicit FROC decision variable being the RGI threshold of the filtered image threshold. Because images are being thresholded to generate these FROC curves, they exhibit the unusual behavior of starting and ending near the $(0, 0)$ point in FROC space. This is because at a very high RGI threshold, no pixels pass the threshold, and one achieves 0 true candidates and 0 false candidates. Also, at a very low threshold, all pixel pass and one is left with a single very large detection. Thus, at very low thresholds, one achieves a sensitivity near 0 and nearly 1 false candidate per image. As is shown in Fig. 5, this technique alone can achieve a sensitivity of 93% with 16 false candidates per image at an RGI threshold of 0.74. A lesion was considered "detected" for this study if the center of mass of a connected region in the thresholded image was within a radiologist's outline of the lesion. The regions returned by this method represent the candidate lesions to be further analyzed by the subsequent feature extraction and pattern classification stages of the computerized mass detection method.

C. Lesion Segmentation

In order to automatically extract features from each candidate lesion, the potential abnormality needs to be segmented from the breast parenchyma background using as input a given seed location as returned by the previous initial detection stage. We have developed and investigated methods of lesion segmentation that involve multiplication of the suspect location (given the seed point) with a constraint function such as a Gaussian and generating a series of lesion-like contours by performing local thresholding on this "constrained" image [44]. The final contour is identified by means of either an RGI-based or a probabilistic method as described elsewhere in the literature [44]. The RGI-based and probabilistic methods outperformed a conventional region growing method when compared with the radiologist's outlined truth. The probabilistic segmentation method tends to be time consuming, so we implemented the RGI-based segmentation in the mass detection method.

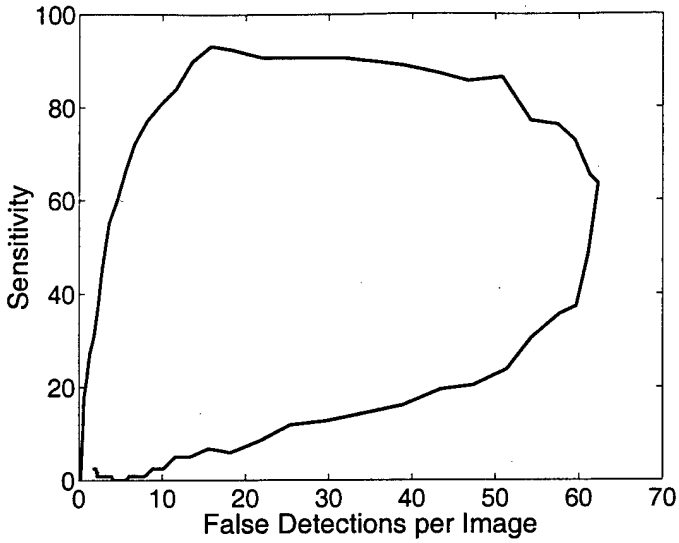


Fig. 5. RGI filtering FROC curves for various minimum size cutoffs using the RGI threshold as the decision variable. Note that there is a large decrease in the false-detection rate when the minimum size cutoff is increased from 0 to 1 without a large decrease in the sensitivity.

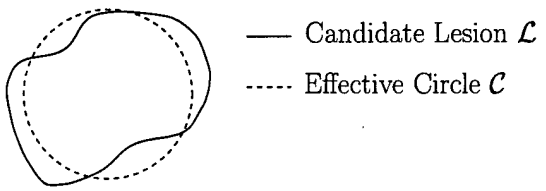


Fig. 6. The sets used to define the geometric-based features.

D. Feature Extraction

The image data and the contour returned by the lesion segmentation stage are employed to extract features for each candidate lesion within the image. These features are subsequently used in the pattern classifier to determine the final classification of each candidate lesion. Radiographically, mass lesions can be characterized by their degree of spiculation, margin definition, shape, density, homogeneity (texture), asymmetry, and so forth. Descriptors of these characteristics may be also grouped as gradient-based features, intensity-based features, and geometric features.

D.1 Geometric Features

Geometric features use the results of the lesion segmentation algorithm to arrive at measures of the size, circularity, irregularity, and compactness [45] of each candidate lesion. The geometric features are defined as:

$$\begin{aligned} \text{Irregularity} &= 1 - \frac{\text{Perim}(\mathcal{C})}{\text{Perim}(\mathcal{L})} \\ \text{Circularity} &= \frac{\text{Area}(\mathcal{L} \cap \mathcal{C})}{\text{Area}(\mathcal{L})} \\ \text{Compactness} &= 4\pi \frac{\text{Area}(\mathcal{L})}{\text{Perim}(\mathcal{L})^2}, \end{aligned}$$

where $\text{Perim}(\cdot)$ is the perimeter operator, $\text{Area}(\cdot)$ is the area operator, $\mathcal{L}$ is the set of lesion pixels, and $\mathcal{C}$ is the effective circle set which is centered at the candidate lesion's center of mass and has an area equal to that of the candidate lesion as illustrated in Fig. 6.

D.2 Gray-level Features

The gray-level features use both the pixel-value information (*i.e.*, the image function $f(x,y)$) as well as the lesion segmentation results (*i.e.*, $\mathcal{L}$) to compute the average gray level within the lesion, the standard deviation of the gray levels within the lesion, the internal contrast

$$IC = \frac{\max\limits_{(x,y)\in\mathcal{L}} f(x,y)}{\min\limits_{(x,y)\in\mathcal{L}} f(x,y)}, \tag{1}$$

and the external contrast

$$EC = \frac{2(AvgI - AvgE)}{AvgI + AvgE}, \tag{2}$$

where $AvgI$ is the average pixel value within the lesion $\mathcal{L}$, and $AvgE$ is the average pixel value within a periphery region outside of the lesion. To determine the periphery neighborhood, a bounding box for each candidate lesion was extended by 3mm on all sides and a smoothed version [46] of the candidate lesion contour was employed to determine which image pixels to exclude (see Fig. 7 for an example periphery region).

D.3 Gradient Features

Gradient-based features are measures such as the average and standard deviation of the gradient strengths within the four neighborhoods [46] (margin, grown region, region-of-

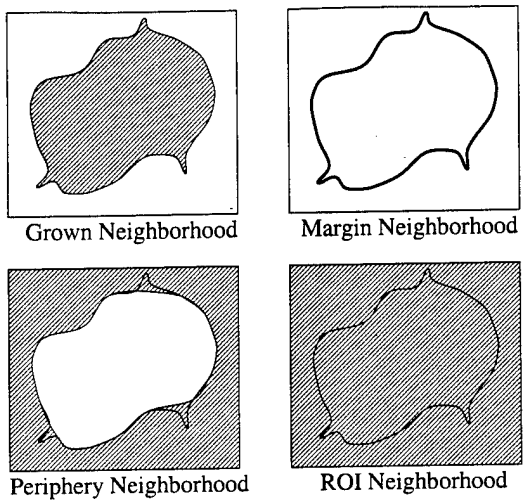


Fig. 7. The four neighborhoods used to calculate gradient-based features.

interest ROI, and periphery) as shown in Fig. 7, the RGI value, and numerous gradient-weighted histogram (GWH) features [46]. To compute the GWH features, one first computes the gradient vectors for each pixel within the neighborhood of interest. Then a histogram is generated of the either the Cartesian angles of the gradients (Cartesian gradient-weighted histograms or CGWHs) or the angles of the gradients relative to the radial direction (radial gradient-weighted histograms or RGWHs). Each entry that is accumulated into these histograms is weighted by the magnitude of the gradients at each angle [46]. Kernel density estimation using a Gaussian kernel with the width determined optimally *via* cross-validation is used to achieve a continuous estimate of the histogram function. Finally, measures such as the full width at half maximum, the minimum histogram value, and the height are computed as illustrated in Fig. 8. GWH analysis is performed on the four neighborhoods (Fig. 7) for each candidate lesion and in both the Cartesian (CGWH) and radial directions (RGWH). GWH features can characterize margin sharpness, spiculation, linearity, as well as other properties of candidate lesions.

In total, 40 features (31 gradient-based, 4 intensity-based, and 5 geometric) are extracted from each candidate lesion site. A subset of these features must be used in the final candidate lesion classification stage. In general, use of all 40 features would require a much larger training database and would severely limit the robustness of the detection

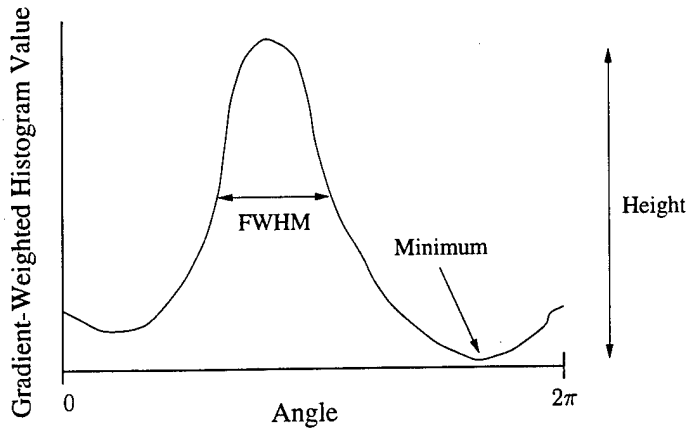


Fig. 8. An illustration of the features extracted from the gradient weighted histograms.

method (*i.e.*, the potential for overtraining would be large). It should be noted that although the pixel size of the digitized mammograms was 100 μm, subsampled images with effective pixel sizes of 500 μm were employed in the initial lesion detection and in the calculation of the feature values.

E. Classification

The task in this stage is to use some of the features described in Section III-D to classify the candidate lesion sites as either malignant lesions or false candidates. We have implemented a Bayesian ANN and a rule-based classifier that we train using a multiobjective approach to classify candidate lesions. These methods will be discussed in detail in the next sections.

IV. Application of the Bayesian ANN

A. Feature Selection

We previously demonstrated [18] the bias that is introduced if one uses the same dataset for both selecting features and determining the parameters of a classifier. If the validation dataset is not employed in the feature selection process, then the performance of the classifier on the independent validation dataset will be depressed. Thus, if the performance on the validation dataset of the classifier using features selected (using the training dataset) by an automated feature selection algorithm is better (and the results are statistically

significant) than that of hand-selected features, then the features selected using the automated technique are "better" than those selected by-hand. It is important to keep small the number of performance comparisons measured on the validation dataset in order to reduce random effects.

We know from our prior simulation studies [17] that in order to avoid overtraining, we can use a maximum of five features in the Bayesian ANN based on our training dataset size. Interestingly, we analyzed the performance of the Bayesian ANN on the training data using various subsets of six features and found that it was susceptible to overtraining as indicated by the the fact that the performance of the Bayesian ANN was not stable with increasing hidden units. Thus, we define the feature selection task as one of selecting the five "best" features to be used as inputs for the Bayesian ANN. Features that based upon similar lesion characteristics were employed because they measure the characteristics in different fashions. For example, two spiculation features are, generally, somewhat correlated but each also contains information that the other feature does not extract and, thus, both features may be useful.

Malignant mass lesions tend to exhibit spiculation, they are more circularly shaped than a typical false candidate, they have a higher density, the pixel values tend to be less variable (*i.e.*, smoother texture), and the margins of lesions are more well-defined than false candidates [47]. Based on this information and the individual performances of the various features in the database as measured on the training dataset, the features characterized in Table I were hand-selected to be used in the Bayesian ANN. The texture feature was not included in the hand-selected feature set because it performed poorly by itself in the task of distinguishing between mass lesions and false candidates.

We also implemented two automated feature selection methods to select features to be used in the Bayesian ANN (Fig. 1). The first was a forward selection method [48] in which the features that improved the performance the most were added in succession. For example, if a contrast measure alone has the best individual Bayesian ANN performance, then contrast is added to the list of selected features. Then, every other feature is looked at in combination with contrast to see which other feature is the most beneficial when used in the Bayesian ANN with contrast. This continues until a specified number of features is

TABLE I

CHARACTERIZATION OF THE FIVE HAND-SELECTED FEATURES IN TERMS OF THE PROPERTIES (*i.e.*, SHAPE, DENSITY, TEXTURE, SPICULATION, AND/OR THE MARGINS) EACH FEATURE IS MEASURING.

Feature	Shape	Density	Texture	Spiculation	Margin
RGI	x				x
Contrast		x			
Margin Strength					x
Margin RGWH FWHM				x	
Margin RGWH Height				x	x

TABLE II

CHARACTERIZATION OF THE FEATURES SELECTED USING A FORWARD SELECTION METHOD.

Feature	Shape	Density	Texture	Spiculation	Margin
Circularity	x				
Margin Strength					x
Contrast		x			
Margin RGWH Height				x	x
Margin CGWH Height	x			x	

selected. Table II shows the features that were automatically selected using this technique and the training dataset. Thirdly, a genetic algorithm (GA) feature selection method [49,50] was implemented in order to select, again, five features to be used in the Bayesian ANN. The five GA-selected features are shown in Table III.

B. Performance

Figure 9 shows the training and validation dataset ROC curves for the three different subsets of five features (Tables I, II, and III) when trained using a Bayesian ANN with 10 hidden units for each. The validation dataset A_z values (Fig. 9(b)) for the hand-selected, forward-selected, and GA-selected features were 0.83, 0.88, and 0.89, respectively, in the task of distinguishing between actual lesions and false candidates. The difference in A_z

TABLE III

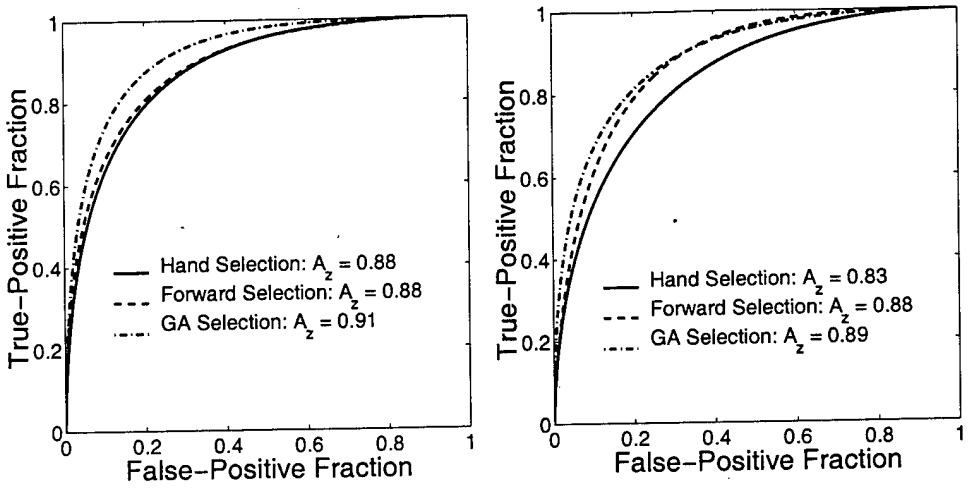
A CHARACTERIZATION OF THE FEATURES SELECTED USING A GENETIC ALGORITHM.

Feature	Shape	Density	Texture	Spiculation	Margin
Circularity	x				
Area	x				
Contrast		x			
Grown RGWH Height				x	x
Grown CGWH Minimum	x				x

between the hand-selected and the forward-selected feature sets on the validation dataset was statistically significant ($p < 0.01$) as it was for the difference in A_z between the hand-selected and GA-selected feature sets on the validation dataset. There was not enough evidence to support a measurable difference in A_z between the forward-selected subset of features and the GA-selected subset of features. It should be noted, however, that the forward selection method is substantially quicker than the GA feature selection method. The differences between the training dataset ROC curves and the validation dataset ROC curves are consistent with the known natural bias that classification systems have [51,52].

C. Resampling Performance

We previously showed [18] that it is unlikely that one will select the optimal subset of features when one has finite data. The various subsets of features described above for use in the Bayesian ANN are likely to be near-optimal but still sub-optimal. Thus, if the database used in this study was repartitioned into different training and validation datasets, different features and different classifier performances would be observered. To study this effect, we repartitioned the original database by-case into ten pairs of randomly selected (by-case) training and validation datasets. The forward selection method was then applied to each of the ten training datasets. Table IV shows the the number of times out of the ten runs of the forward feature selection method that each feature from Table II was selected. We also computed the average training dataset A_z and validation dataset A_z values which were 0.91 ± 0.01 and 0.87 ± 0.02, respectively. The average ROC



(a) (b)

Fig. 9. (a) Training and (b) validation performances of the Bayesian ANN with three different subsets of 5 features. The classification task was to distinguish between actual lesions and false candidates.

curves (obtained by averaging the a and b ROC curve parameters [53]) are shown in Fig. 10. It is clear from Table IV that a couple of the features are frequently selected by the forward selection method and are, thus, important for the classification of the lesions. Because other features are less often selected and because the performance of the Bayesian ANN is consistent, these features are "replaceable" by other features that measure similar characteristics.

V. Application of the Multiobjective Approach

The Bayesian ANN approximates the ideal observer but typically requires numerous parameters to do so [17]. It is often desirable to use a classifier with simple and understandable rules. Because simple rules (*i.e.*, few classifier parameters) are employed, it is possible to incorporate more features into the classification system than would be possible with an ANN. We implemented a simple thresholding rule-based classifier in which features are sequentially thresholded to determine the class of the candidate lesions. An example of a thresholding rule-based classifier would be to call a candidate lesion a malignant lesion if and only if the circularity is greater than 0.5, the contrast is greater than

TABLE IV

THE NUMBER OF TIMES OUT OF THE TEN DIFFERENT FORWARD SELECTION RUNS THAT VARIOUS FEATURES WERE SELECTED. THE FEATURES SHOWN IN THIS TABLE ARE THE ONE SELECTED BY THE FORWARD FEATURE SELECTION METHOD ON THE ORIGINAL PARTITION OF THE DATASET (*i.e.*, TABLE II).

Feature	Number of times selected out of 10
Circularity	4
Margin Strength	2
Contrast	7
Margin RGWH Height	9
Margin CGWH Height	1

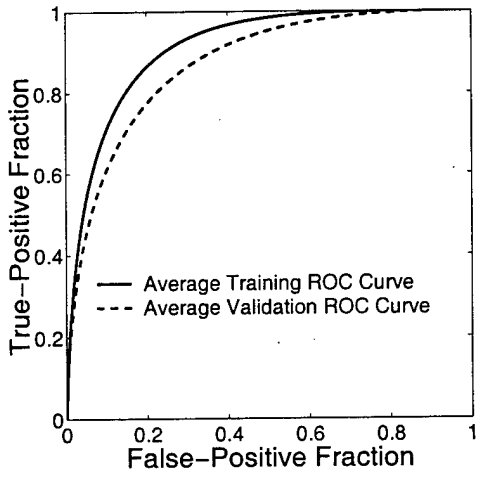


Fig. 10. The average Bayesian ANN ROC curves for the 10 repartitions of the entire dataset. The classification task was to distinguish between actual lesions and false candidates in mammography.

0.2, and the margin sharpness is above 1.2. If any of these conditions are not met then the candidate lesion is considered a false detection. We used the Niched Pareto genetic algorithm (NP-GA) described in [19] to design a thresholding rule-based classifier for the computerized detection of malignant, mass lesions in mammography. Results on both the training and validation datasets are presented.

A. Feature Selection

Because the classifier we are implementing requires only thresholding rules, feature selection is a simpler task. One must choose features that individually perform well (*i.e.*, have a high A_z) and are not highly correlated with one another (*i.e.*, do not provide redundant information). Because fewer parameters are used with this classifier we can likely use more input features than was possible with the Bayesian ANN. We selected eight features to be used in the classifier. These eight features selected were done so based on expert knowledge (*i.e.*, based on what radiologists' use to make decisions), individual performance (*i.e.*, A_z), and correlations with other features. The characteristics of these eight features are listed in Table V. The maximum correlation among these features was 0.74. Note that this set of features is a superset of the features selected by the forward-selection method shown in Table II. Other similar subsets of 8 features were also evaluated but the results did not seem to vary greatly. Subsets of fewer than eight features were also evaluated and found to perform, on average, worse than did the eight features. Subsets of more than eight features were evaluated and found to have similar performances as well, thus there did not seem to be any benefit in adding additional features beyond eight.

B. NP-GA Performance

Figure 11 shows the performance of the NP-GA trained rule-based classifier on the training and validation datasets. The A_z of the rule-based classifier was 0.85 on the training dataset and 0.80 on the validation dataset. The substantial difference in performance between the training and validation datasets may indicate that the NP-GA is fine-tuning the thresholds to such an extent that they are not generalizable when applied to an independent dataset. However, the number of parameters used in the rule-based classifier was 8 as compared with the 71 used in the Bayesian ANN (*i.e.*, a Bayesian ANN with 5

TABLE V

A CHARACTERIZATION OF THE FEATURES SELECTED FOR USE IN THE NP-GA.

Feature	Shape	Density	Texture	Spiculation	Margin
Circularity	x				
RGI	x				x
Contrast		x			
Pixel Variation			x		
Margin Strength					x
Margin RGWH FWHM				x	
Margin RGWH Height				x	x
Margin CGWH Height	x			x	

input features, 10 hidden nodes and 1 output node has a total of 71 parameters) which was able to achieve an A_z of 0.89 on the validation dataset. Thus, it is interesting to note the high performance of the rule-based classifier trained using the NP-GA and using very few parameters.

In order to fit the ROC curves shown in Fig. 11, one must generate decision variable data to be input into the ROC curve fitting software. To accomplish this, the NP-GA returned ROC operating points, and the *a priori* knowledge of the numbers of true and false observations in the dataset are used to generate mock decision variable data which is input into the ROC software [53]. If the ROC curve returned by the NP-GA has N operating points (including the $(0,0)$ and $(1,1)$ points in ROC space), then one has $N-1$ bins of decision variable data to generate where values of the decision variable data are $1, 2, \ldots, N-1$. One fills in these bins of data such that if N thresholds are applied between the bins, then the sensitivity and specificity values of the ROC curve are achieved. For example, if one has 100 true and 100 false detections in the dataset and the 4th and 5th adjacent operating points in the ROC curve have sensitivities of 90% and 85%, respectively, then one generates 5 ($90\% - 85\% = 5\%$ of 100) mock true-positive decision variable data with values of 4 (the 4th operating point). If, the 6th operating point has a sensitivity of 70%, then one generates 15 ($85\% - 70\% = 15\%$ of 100) mock true-positive decision

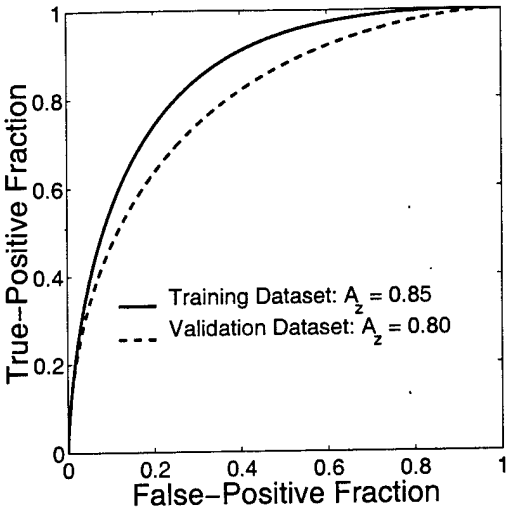


Fig. 11. The performance of the NP-GA trained rule-based classifier on the training and validation datasets in the task of distinguishing between actual lesions and false candidates.

variable data with values of 5 (the 5th operating point). A similar strategy is taken with the false-positive decision variable data as well.

C. Resampling Performance

An evaluation of the performance of the rule-based classifier trained using the NP-GA using different dataset partitions was performed. In this study, the features used in the classifier were fixed (*i.e.*, the features shown in Table V) and the NP-GA was rerun 10 times on each training dataset and then tested on each validation dataset. The average training dataset A_z was 0.87 ± 0.02, and the average validation dataset A_z was 0.81 ± 0.02. The average ROC curves (obtained by averaging the 10 a and b ROC curve parameters [53]) are shown in Fig. 12.

VI. Discussion

The Bayesian ANN can achieve an A_z of 0.89 in the task of distinguishing between malignant mass lesions and false candidates as was shown in Fig. 9(b). Using more features with the rule-based classifier trained using the NP-GA, we achieved an A_z of 0.80 in the task of distinguishing between malignant mass lesions and false candidates. While it is generally not true that a Bayesian ANN with fewer features will always outperform an

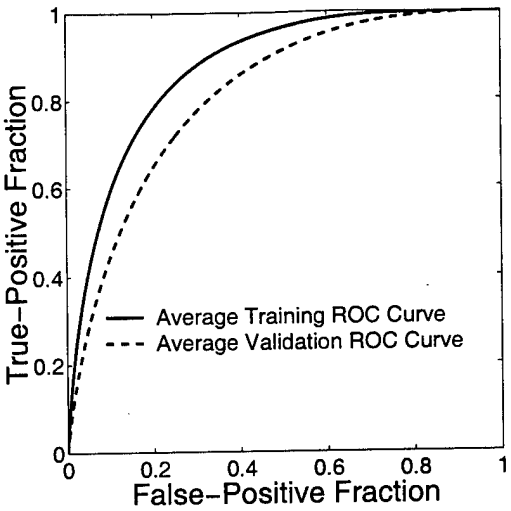


Fig. 12. The average training dataset and validation dataset ROC curves for 10 random partitions of the entire dataset into training and validation datasets. The classification task was to distinguish between actual lesions and false candidates in mammography.

NP-GA trained rule-based classifier, it is true for the classification task discussed in this paper.

It is apparent from Fig. 9 that the Bayesian ANN is not becoming too specific to the training dataset and the results are generalizable to the validation dataset and other independent datasets. However, from Fig. 11, one can see that the NP-GA rule-based classifier is becoming too specific to the training dataset. The NP-GA is designed to produce the best possible ROC curve that the given classifier (a rule-based classifier in this case) can achieve on the given training dataset. Thus the threshold values can become too specific to the training dataset. This causes the NP-GA trained rule-based classifier to perform poorly on the validation dataset. Although not shown in this work, different numbers of features and different subsets of features were applied to the NP-GA training with similar results.

Thus far, we have shown the performance of only the classifier in the task of distinguishing between malignant mass lesions and false candidates. In order to evaluate the overall mass detection method, FROC analysis [54–56] is employed. The Bayesian ANN was chosen as the final classifier with the input features being those selected using the

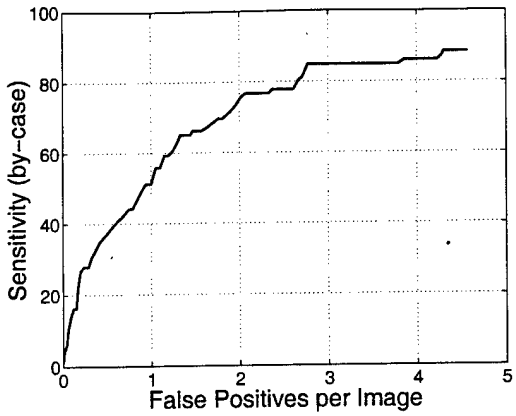


Fig. 13. The performance of the overall mass detection method in the task of distinguishing between actual lesions and false candidates using the Bayesian ANN with 5 input features as the classifier.

forward selection method (Table II). If a candidate lesion was classified by the Bayesian ANN as being a malignant lesion and that candidate lesion overlapped at all with a radiologist's outline of the true lesions, then that lesion was considered to be detected by the mass detection method. We plot our FROC curve using the by-case sensitivity, *i.e.*, a lesion needed only be located by the computer in one image to be considered "detected". Also, if there are more than 5 detections in a single image, then only the 5 detections with the highest ANN output values are used so as to limit the total number of detections in a single image. Figure 13 shows the by-case FROC curve for the mass detection method using the Bayesian ANN as the classifier. Using the Bayesian ANN, we were able to achieve a sensitivity of 75% at 2 false positives per image on the validation dataset.

The performance of the mass detection method characterized in Fig. 13 is affected not only by the pattern classification stage but also by the performance of the initial detection method, the ability of the segmentation algorithm to accurately segment potential abnormalities from surrounding parenchyma given the initial detection information, and the ability of the feature extraction stage to extract useful information for the classifier. Thus, to improve the performance of this mass detection method one can focus their attention on several aspects of the method. First, one could improve upon the initial detection algorithm and design a method that returns fewer false candidates than RGI filtering. One could also extract more useful features from candidate lesion sites or provide extra

features using previous mammograms or different views of the same breast to provide "better" information to the classifier. Finally, one can use larger database sizes and have the Bayesian ANN approximate the ideal observer with more features (*i.e.*, information) used as inputs.

VII. Conclusions

We designed pattern classifiers to distinguish between malignant mass lesions and false candidates in the computerized detection of mass lesions in mammography using both Bayesian ANNs and the multiobjective training approach. Both methods were found to be practically useful in distinguishing between malignant mass lesions and false detections. However, there is a need for further improvements in the performance of the overall technique if it is be used in a clinical setting. It has become readily apparent that one must have large databases to properly train the classifiers using as much useful information as possible. With more useful information extracted and larger training databases, the performance of the overall mass detection algorithm is expected to improve.

Pattern classifiers are widely used in many computerized analysis methods for a wide variety of imaging modalities. Therefore, the methods of pattern classification presented here can be applied to different computerized analysis methods. For example, the multiobjective approach has been successfully applied to the computerized detection of microcalcifications in mammograms by Anastasio *et al.* [57]. Bayesian ANNs have recently been applied to the analysis of trabecular bone patterns for determining the risk of fracture using texture features extracted from conventional radiographs [58], and to the characterization of mass lesions as benign or malignant in small-field digital mammography systems [59].

Acknowledgments

The authors thank Dr. Carl J. Vyborny and Dr. Ulrich Bick from The University of Chicago for their database collection and characterization efforts as well as Dr. Kathy Yao from Northwestern University for her database collection efforts. This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202) and USPHS grants RR11459 and T32 CA09649.

Maryellen L. Giger is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

REFERENCES

[1] *Cancer Facts and Figures-1998.* New York: American Cancer Society, 1998.

[2] *Surveillance, Epidemiology and End Results Program. Statistics Review.* Bethesda, MD: National Cancer Institute, 1992.

[3] L. W. Bassett and R. H. Gold, *Breast cancer detection: Mammography and other methods in breast imaging.* New York: Grune and Stratton, 1987.

[4] C. J. Baines, A. B. Miller, and C. Wall, "Sensitivity and specificity of first screen mammography in the Canadian national breast screening study. A preliminary report from five centers," *Radiology*, vol. 160, pp. 295–298, 1986.

[5] S. R. Pollei, F. A. Mettler, S. A. Bartow, G. Moradian, and M. Moskowitz, "Occult breast cancer. Prevalence and radiographic detectability," *Radiology*, vol. 16, pp. 459–462, 1987.

[6] I. Andersson, "What can we learn from interval carcinomas?," *Recent Results in Cancer Research*, vol. 90, pp. 191–193, 1984.

[7] J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancers missed by mammography," *American Journal of Roentgenology*, vol. 132, pp. 737–739, 1979.

[8] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," *New England Journal of Medicine*, vol. 331, pp. 1493–1499, 1994.

[9] J. R. Buchanann, J. S. Spratt, and L. S. Heuser, "Tumor growth, doubling times, and the inability of the radiologist to diagnose certain cancers," *Radiologic Clinics of North America*, vol. 21, p. 115, 1983.

[10] T. Holland, M. Mrvunac, J. H. C. L. Hendricks, and B. Bekker, "So-called interval cancers of the breast. Pathologic and radiographic analysis," *Cancer*, vol. 49, pp. 2527–2533, 1982.

[11] W. A. Murphy Jr, J. M. Destouet, and B. S. Monsees, "Professional quality assurance for mammography screening programs," *Radiology*, vol. 175, pp. 319–320, 1990.

[12] R. E. Bird, "Professional quality assurance for mammography screening programs," *Radiology*, vol. 177, p. 587, 1990.

[13] R. J. Brenner, "Medicolegal aspects of breast imaging: Variable standards of care relating to different types of practice," *American Journal of Roetgenology*, vol. 156, pp. 719–723, 1991.

[14] L. Tabar, G. Fagerberg, S. W. Duffy, N. E. Day, A. Gad, and O. Grontoft, "Update of the Swedish two-country program of mammographic screening for breast cancer," *Radiol Clin North Am*, vol. 30, pp. 187–210, 1992.

[15] E. L. Thurfjell, K. A. Lernevall, and A. A. Taube, "Benefit of independent double reading in a population-based mammography screening program," *Radiology*, vol. 191, pp. 241–244, 1994.

[16] C. J. Vyborny and M. L. Giger, "Computer vision and artificial intelligence in mammography," *American Journal of Roentgenology*, vol. 162, pp. 699–708, 1994.

[17] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Transactions on Medical Imaging*, 2000 (in review).

[18] M. A. Kupinski and M. L. Giger, "Feature selection with limited datasets," *Medical Physics*, vol. 26, pp. 2176–2182, 1999.

[19] M. A. Kupinski and M. A. Anastasio, "Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 675–685, 1999.

[20] B. Zheng, Y.-H. Chang, and D. Gur, "On the reporting of mass contrast in CAD research," *Medical Physics*, vol. 23, no. 12, pp. 2007–2009, 1996.

[21] Y. Wu, K. Doi, M. L. Giger, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," *Medical Physics*, vol. 19, pp. 555–560, 1992.

[22] N. Asada, K. Doi, H. MacMahon, S. M. Montner, M. L. Giger, C. Abe, and Y. Wu, "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: Pilot study," *Radiology*, vol. 177, pp. 857–860, 1990.

[23] H.-P. Chan, S.-C. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Medical Physics*, vol. 22, pp. 1555–1567, 1995.

[24] J. Y. Lo, J. A. Baker, P. J. Kornguth, and J. C. E. Floyd, "Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features," *Academic Radiology*, vol. 2, pp. 841–850, 1995.

[25] D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Medical Physics*, vol. 22, pp. 1501–1513, 1995.

[26] H. D. Li, M. Kallergi, L. P. Clarke, and V. K. Jain, "Markov random field for tumor detection in digital mammography," *IEEE Transactions on Medical Imaging*, vol. 14, no. 3, pp. 565–576, 1995.

[27] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.

[28] N. Karssemeijer, "Recognition of stellate lesions in digital mammograms," in *Digital Mammography* (A. C. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns, eds.), Elsevier, 1994.

[29] T. Parr, R. Zwiggelaar, S. Astley, C. Boggis, and C. Taylor, "Comparison of methods for combining evidence for spiculated lesions," in *Digital Mammography*, (Dordrecht, The Netherlands), Kluwer Academic Publishers, 1998.

[30] H. Kobatake, H. Takeo, and S. Nawano, "Tumor detection system for full-digital mammography," in *Digital Mammography*, (Dordrecht, The Netherlands), Kluwer Academic Publishers, 1998.

[31] U. Bick, M. L. Giger, R. A. Schmidt, and K. Doi, "A new single-image method for computer-aided detection of small mammographic masses," in *CAR '95: International Symposium on Computer and Communication Systems for Image Guided Diagnosis and Therapy*, pp. 357–363, 1995.

[32] H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," *Radiology*, vol. 212, pp. 817–827, 1999.

[33] B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics*, vol. 25, pp. 516–526, 1998.

[34] M. L. Giger, "Future of breast imaging. Computer-aided diagnosis," in *AAPM/RSNA Categorical Course on the Technical Aspects of Breast Imaging* (A. Haus and M. Yaffe, eds.), pp. 257–270, AAPM/RSNA, 1992.

[35] K. Doi, M. L. Giger, R. M. Nishikawa, K. R. Hoffmann, H. MacMahon, and R. A. Schmidt, "Potential usefulness of digital imaging in clinical diagnostic radiology: Computer-aided diagnosis," *Journal of Digital Imaging*, vol. 8, pp. 2–7, 1995.

[36] M. L. Giger, "Current issues in CAD for mammography," in *Digital Mammography* (K. Doi, ed.), International Congress Series, pp. 53–59, Elsevier, 1996.

[37] M. L. Giger, "Computer-aided diagnosis," in *AAPM/RSNA Categorical Course in Diagnostic Radiology Physics: Physical Aspects of Breast Imaging–Current and Future Considerations* (A. Haus and M. Yaffe, eds.), pp. 249–272, AAPM/RSNA, 1999.

[38] C. J. Vyborny, "Can computers help radiologists read mammograms?," *Radiology*, vol. 191, pp. 315–317, 1994.

[39] A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns, *Digital Mammography*. Amsterdam: Elsevier, 1994.

[40] K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, *Digital Mammography '96*. Amsterdam: Elsevier, 1996.

[41] N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, *Digital Mammography 1998*. Kluwer Academic Publishers, 1998.

[42] K. Doi, H. MacMahon, M. L. Giger, and K. R. Hoffmann, eds., *Computer-Aided Diagnosis in Medical Imaging*. Amsterdam: Elsevier, 1999.

[43] M. A. Kupinski and M. L. Giger, "Detection of mass lesions in mammography using radial gradient index filtering," *IEEE Transactions on Medical Imaging*, 2000 (under review).

[44] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 510–517, 1998.

[45] T. Matsumoto, H. Yoshimura, K. Doi, M. L. Giger, A. Kano, H. MacMahon, K. Abe, and S. M. Montner, "Image feature analysis of false-positive diagnoses produced by automated detection of lung nodules," *Investigative Radiology*, vol. 27, pp. 587–597, 1992.

[46] Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, and P. Lu, "Analysis of spiculation in the computerized classification of mammographic masses," *Medical Physics*, vol. 22, pp. 1569–1579, 1995.

[47] C. J. Vyborny and R. A. Schmidt, "Mammography as a radiographic examination: An overview," *Radio-Graphics*, vol. 9, pp. 723–764, 1989.

[48] M. A. Kupinski and M. L. Giger, "Feature selection and classifiers for the computerized detection of mass lesions in digital mammography," in *Proceedings of the 1997 International Conference on Neural Networks (ICNN '97)*, (Houston, TX), pp. 2460–2463, IEEE/ICNN, June 9-12 1997.

[49] M. A. Kupinski, M. L. Giger, and K. Doi, "Optimization of neural network inputs with genetic algorithms," in *Digital Mammography* (K. Doi, ed.), International Congress Series, pp. 401–404, Elsevier, 1996.

[50] M. A. Kupinski and M. L. Giger, "Genetic algorithm feature selection in CAD." Undergraduate Thesis/Trinity University, San Antonio, TX, 1995.

[51] H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Medical Physics*, vol. 26, no. 12, pp. 2654–2668, 1999.

[52] R. F. Wagner, H.-P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling

plans: Applications to linear classifiers in computer-aided diagnosis," in *SPIE Medical Imaging 1997*, vol. 3034, pp. 467–477, 1997.

[53] C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statistics In Medicine*, vol. 17, pp. 1033–1053, 1998.

[54] J. P. Egan, G. Z. Greenberg, and A. I. Schulman, "Operating characteristics, signal detectability, and the method of free response," *Journal of the Acoustical Society of America*, vol. 33, pp. 993–1007, 1961.

[55] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free-response approach to the measurement and characterization of radiographic-observer performance," *Journal of Applied Photographic Engineering*, vol. 4, pp. 166–171, 1978.

[56] D. P. Chakraborty and L. H. L. Winder, "Free-response methodology: Alternate analysis and a new observer-performance experiment," *Radiology*, vol. 174, pp. 873–881, 1990.

[57] M. A. Anastasio, M. A. Kupinski, and R. M. Nishikawa, "Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 1089–1093, 1998.

[58] M. Chinander, M. L. Giger, J. Favus, C. Jiang, M. Favus, and M. A. Kupinski, "Application of Bayesian artificial neural networks to the assessment of bone strength," in *World Congress on Medical Physics and Biomedical Engineering*, AAPM/EMBS, 2000 (to be published).

[59] M. L. Giger, M. Maloney, L. Venta, Z. Huo, C. J. Vyborny, and M. A. Kupinski, "Computerized classification of lesions on digital mammography," in *International Workshop on Digital Mammography 2000*, IWDM, 2000 (to be published).

Feature selection with limited datasets

Matthew A. Kupinski and Maryellen L. Giger
Kurt Rossmann Laboratories, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Avenue, Chicago, Illinois 60637

(Received 26 January 1999; accepted for publication 19 July 1999)

Computer-aided diagnosis has the potential of increasing diagnostic accuracy by providing a second reading to radiologists. In many computerized schemes, numerous features can be extracted to describe suspect image regions. A subset of these features is then employed in a data classifier to determine whether the suspect region is abnormal or normal. Different subsets of features will, in general, result in different classification performances. A feature selection method is often used to determine an "optimal" subset of features to use with a particular classifier. A classifier performance measure (such as the area under the receiver operating characteristic curve) must be incorporated into this feature selection process. With limited datasets, however, there is a distribution in the classifier performance measure for a given classifier and subset of features. In this paper, we investigate the variation in the selected subset of "optimal" features as compared with the true optimal subset of features caused by this distribution of classifier performance. We consider examples in which the probability that the optimal subset of features is selected can be analytically computed. We show the dependence of this probability on the dataset sample size, the total number of features from which to select, the number of features selected, and the performance of the true optimal subset. Once a subset of features has been selected, the parameters of the data classifier must be determined. We show that, with limited datasets and/or a large number of features from which to choose, bias is introduced if the classifier parameters are determined using the same data that were employed to select the "optimal" subset of features. [S0094-2405(99)01010-X]

Key words: feature selection, classification, computer-aided diagnosis

I. INTRODUCTION

Computerized diagnosis schemes have the potential of increasing diagnostic accuracy in radiological imaging.[1–4] In computerized schemes, features characterizing suspicious image regions are extracted and input to a data classifier to predict pathology.[5–7] Different combinations of features will, in general, yield different classification performances. In addition, relatively few features should be used in order for the classifier to remain robust. Thus, one is faced with the task of selecting a useful and limited subset of features from many available features.

In automated feature selection, a computer algorithm determines the subset of features that will result in the best classification performance. Jain *et al.*[8] review the different types of feature selection methods, which they broadly categorized into deterministic methods such as stepwise feature selection,[9] stochastic methods such as genetic algorithm feature selection,[7,10,11] and optimal methods that are usually prohibitively time consuming such as an exhaustive search of all possible feature subset combinations. Feature selection is becoming a vital step in many computerized schemes due to the large number of features that can be extracted from suspicious image regions.[7,10] Sample sizes, however, are often limited for these studies due to the nature of the problems; for example, abnormal cases are difficult to obtain in many diagnostic procedures.

In this paper, we study the effect of limited sample sizes and large total numbers of features on the ability of a feature selector to select the optimal subset of features. We also study the bias introduced if one uses the same data in selecting features and in determining the parameters of the classifier. Section II contains an introduction to classifier feature selection and introduces the performance measures employed in this work. In Sec. III the difficulty of feature selection with limited datasets is demonstrated using analytical calculations of a simple feature selection problem. Section IV discusses the bias introduced when the same dataset is employed to determine the subset of features and the parameters of the classifier. Finally, Sec. V includes a discussion of the implications of this work to other feature selection methods.

II. BACKGROUND

A. Feature selection and classifiers

A binary classifier takes an observation of feature values and determines whether the observation belongs to either the abnormal class π_a or the normal class π_n. In order to achieve optimal performance, both a set of optimal features and the optimal parameters (structure) of the classifier need to be determined. The N features values corresponding to a particular observation that is to be classified can be expressed by a N-dimensional random vector $\vec{\mathbf{x}}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$ where bold symbols denote random variables. We label the N features making up this random vector

 0094-2405/99/26(10)/2176/7/$15.00

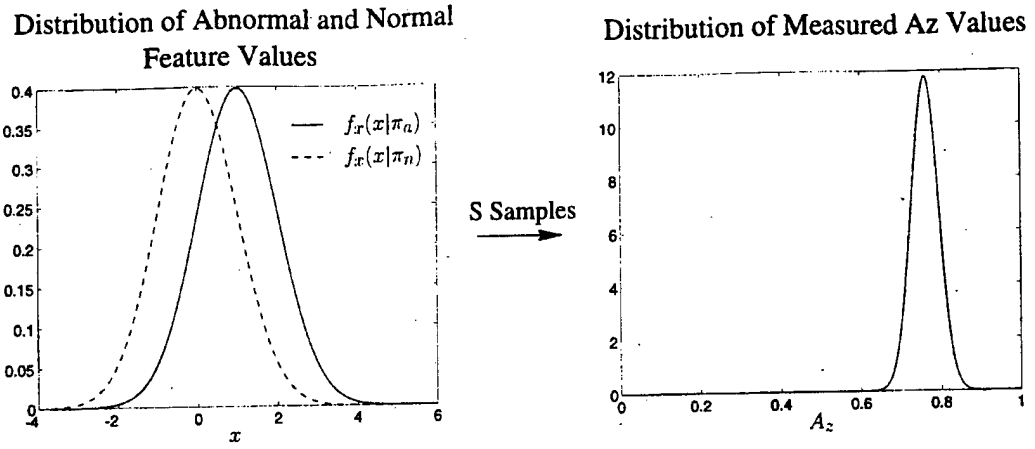


FIG. 1. When limited data are present, there is randomness associated with the measurement of the performance of a classifier. This randomness can be modeled. From knowledge of the underlying density functions of the data (left graph) and the number of samples in each class, we can estimate the distribution of measured A_z values (right graph). The density functions of the data are represented by $f_x(x|\pi_a)$ for the abnormal class and $f_x(x|\pi_n)$ for the normal class.

by the set $\mathcal{F}$. A sample of data called the training dataset, along with its corresponding truth information (π_a or π_b), is used to determine the parameters of the classifier or to select a set of features. Throughout this paper, we use the notation that there are S_a abnormal cases and S_n normal cases in the training dataset where $S_a+S_n=S$. Often an independent dataset with a total of $S'=S'_a+S'_n$ samples is employed for testing.

It is well known that determining the parameters of a classifier with many features can be detrimental due to the "curse of dimensionality." It is therefore often necessary to select a subset of features to use for classification. For any given subset of features $\mathcal{Y}\subseteq\mathcal{F}$, we can define a classifier performance measure $g(\mathcal{Y})$ which measures the performance of the subset of features $\mathcal{Y}$ in the task of distinguishing between the class π_a and the class π_n. The task of feature selection, as proposed by Jain *et al.*,[8] is to select a subset of n features denoted by the set $\mathcal{Z}$ from the entire set of features $\mathcal{F}$ such that

$$g(\mathcal{Z})=\max_{\mathcal{Y}\subseteq\mathcal{F}|\mathcal{Y}|=n} g(\mathcal{Y}) \tag{1}$$

where $|\mathcal{Y}|$ is the number of elements in the set $\mathcal{Y}$. We assume, without loss of generality, that we are trying to maximize the function $g(\cdot)$. In this work we concern ourselves with the task of selecting the optimal subset of n features and not with the task of selecting both the optimal subset and the number of features within that subset.

B. Performance measures

Receiver operating characteristic (ROC) analysis[12,13] is an accepted method for evaluating the performance of classifiers. The area under the maximum likelihood estimate of the ROC curve, A_z,[14] is commonly used as a single performance measure to characterize the performance of a classifier despite the fact that there are some drawbacks to this or any scalar classifier performance measure.[13] In this work we investigate the effect of the distribution of A_z measurements due to finite sample size on the ability to select an optimal subset of features. In theory, however, there will always be a distribution associated with any measure that employs limited data. The results presented in this paper can thus be generalized to different performance measures such as the partial area index[15] or other traditional performance measures.[16] Attention is focused on A_z, however, because it is commonly used within the diagnostic community to select features and to characterize the performance of classifiers.

From knowledge of the theoretical A_z and the numbers of observations in each class (S_a and S_n), we can estimate the distribution of the random variable $\mathbf{A_z}$ of measured A_z values using a Gaussian centered at the theoretical A_z with an estimate of the standard deviation given as

$$\sigma=\sqrt{\frac{1}{S_aS_n}\left(A_z(1-A_z)+(S_a-1)\left(\frac{A_z}{2-A_z}-A_z^2\right)+(S_n-1)\left(\frac{2A_z^2}{1+A_z}-A_z^2\right)\right)}. \tag{2}$$

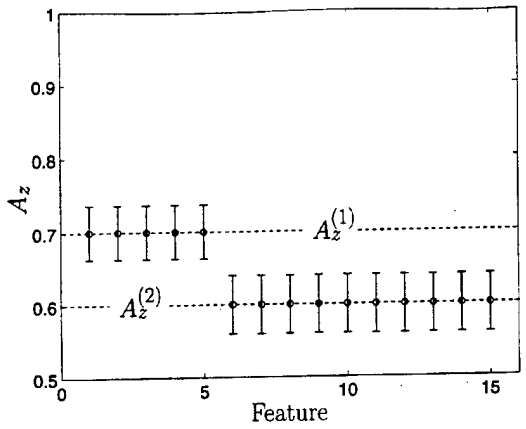


FIG. 2. An illustration of a feature selection problem. In this example, one is selecting a total of n features from a total of $N=15$ independent features where $r=5$ features have a theoretical A_z value of $A_z^{(1)}=0.7$ and $N-r$ features have a theoretical A_z value of $A_z^{(2)}=0.6$. We will consider the situation in which $n \leq r$. The distribution of $\mathbf{A_z}$ measurements is signified by the error bars on each feature's A_z value.

Equation (2) was derived by Hanley *et al.*[17] using the statistical properties of the Wilcoxon statistic to predict the statistical properties of the area under an ROC curve. Figure 1 illustrates the distribution of $\mathbf{A_z}$ measurements for a single feature denoted by the random variable $\mathbf{x}$. From knowledge of the underlying distribution of the data [Fig. 1(a)], we can estimate the distribution of the measured $\mathbf{A_z}$ values [Fig. 1(b)]. Both the theoretical true A_z and the sample sizes (S_a and S_n) determine the distribution of measured $\mathbf{A_z}$ values. The Gaussian model for the density of the measured $\mathbf{A_z}$ fails when the theoretical A_z value is greater than approximately 0.95. The theoretical A_z values for all the simulation studies performed in this paper are below 0.95. We will throughout this paper use the notation that an estimate of the area under the ROC curve is given by an $\mathbf{A_z}$ and the theoretical true area is given by A_z.

III. SELECTING A SUBSET OF FEATURES

Let us consider the following ideal situation: We have a total of N independent features with the first r features having a theoretical individual A_z value of $A_z^{(1)}$ and the remaining $N-r$ features having a theoretical individual A_z of $A_z^{(2)}$ where $A_z^{(1)}>A_z^{(2)}$. Because the features in this situation are independent, we conclude that the random variables denoting the N measured $\mathbf{A_z}$ values are also independent with density functions given by $f_1(A_z)$ for the r features with a theoretical A_z values of $A_z^{(1)}$ and $f_2(A_z)$ for the $N-r$ features with a theoretical A_z value of $A_z^{(2)}$. Similarly, the distribution functions are given by $F_1(A_z)$ for the r features with a theoretical A_z values of $A_z^{(1)}$ and $F_2(A_z)$ for the $N-r$ features with a theoretical A_z value of $A_z^{(2)}$. Figure 2 illustrates such an example where the error bars represent the standard deviations of the measured $\mathbf{A_z}$ values of each feature [Eq. (2)].

The task in this situation is to select the n features (where $n \leq r$) that have the largest measured $\mathbf{A_z}$ values. Because, however, the measured $\mathbf{A_z}$ values have a distribution associated with them, there is a measurable probability that one or more of the "worse" features (those features with a theoretical A_z value of $A_z^{(2)}<A_z^{(1)}$) will be selected. Using order statistics,[18,19] we can compute the probability that an optimal subset of features will be selected:

$$P=\frac{r!}{(n-1)!(r-n)!}\int_0^1 dA_z f_1(A_z) \times(1-F_1(A_z))^{n-1}F_1(A_z)^{r-n}F_2(A_z)^{N-r}, \quad (3)$$

where the integration is from 0 to 1 because A_z values are bound between 0 and 1. A derivation of Eq. (3) is given in the Appendix. In theory, the probability in the situation where each independent feature has a different theoretical A_z value could be computed, but it is computationally impractical.

Figure 3(a) plots the probability of an optimal subset of 4 features being selected as a function of the total number of features to select from N [see Eqn. (3)]. In this plot the total number of features with theoretical $A_z=A_z^{(1)}$ was 4, $A_z^{(1)}$ was set at 0.70, and $A_z^{(2)}$ was fixed at 0.60. The sample size S was also varied from 100 to 1000 where there were equal numbers of abnormal and normal observations, i.e., $S_a=S_n=S/2$. As Fig. 3(a) shows, with small sample sizes the probability of selecting an optimal subset of features drops quickly as the total number of features N increases. Figure 3(b) shows similar plots ($n=4$, $r=4$) but with higher theoretical A_z values, i.e., $A_z^{(1)}=0.8$ and $A_z^{(2)}=0.7$. The difference in theoretical A_z values ($A_z^{(1)}-A_z^{(2)}$) is the same in Figs. 3(a) and 3(b), but as these two figures show, the probability of selecting an optimal subset of features depends on the theoretical A_z values and not just the difference between the theoretical A_z's of the "good" and "bad" features. In this case, the probability of selecting an optimal subset of features decreases as the theoretical A_z's decrease.

Figure 4 shows a plot analogous to Fig. 3(a), except that there are now a total of 10 features with a theoretical A_z value of $A_z^{(1)}$, i.e., $r=10$ instead of 4. The number of features to select, however, is still 4 so we are analyzing the probability that the 4 selected features will be a subset of the 10 actually better features. As Figs. 3 and 4 illustrate, when there are more "good" features in the population of features ($r=10$ instead of $r=4$), the probability of selecting an optimal subset tends to increase.

Figures 3 and 4 show the effect of different theoretical A_z values on the probability of selecting an optimal subset of features. Figure 5 demonstrates this effect in more detail by fixing the sample size at $S=100$ and plotting the probability of selecting an optimal subset for various combinations of $A_z^{(1)}$ and $A_z^{(2)}$. As one would expect, when the difference between the "good" and "bad" features is large, it is easier to select an optimal subset than when the difference between the two classes of features is small.

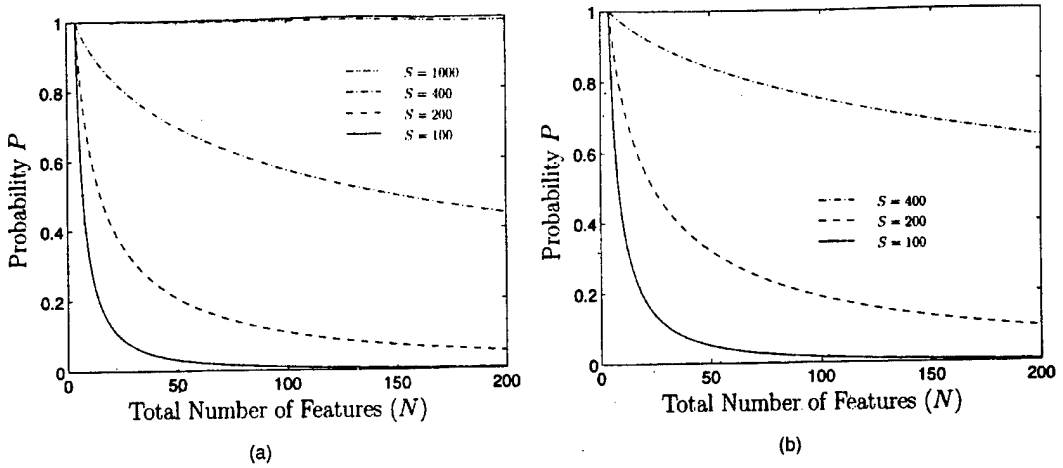


FIG. 3. A plot of the probability of selecting an optimal subset of $n=4$ features from a total of N features. For (a), there are a total of $r=4$ features with a theoretical A_z value of $A_z^{(1)}=0.7$ and $N-r$ features with a theoretical A_z value of $A_z^{(2)}=0.6$. For (b), there are a total of $r=4$ features with a theoretical A_z value of $A_z^{(1)}=0.8$ and $N-r$ features with a theoretical A_z value of $A_z^{(2)}=0.7$. The probability of selecting an optimal subset of features is also plotted for various sample sizes S. The difference in A_z values is 0.1 for both (a) and (b) and yet the probability of selection the optimal subset is higher for (b) than it is for (a). The probability of selecting the optimal subset of features depends of the theoretical A_z values and not just the difference in A_z value between the better and worse features.

IV. CLASSIFICATION WITH SELECTED FEATURES

The feature selection method analyzed in the preceding section selects the n features with the largest measured $\mathbf{A_z}$ values. As illustrated, under certain realistic conditions (small sample sizes and large total numbers of features), it becomes unlikely that this method will select an optimal subset of features. The problems under these conditions, however, are two-fold because it is not only difficult to select an optimal subset (as shown earlier), but data used to classify the features once a subset has been selected will often not be representative of the underlying density functions because feature selection methods, in general, aim to select features with high measured $\mathbf{A_z}$ values. In essence if a poor feature (theoretically poor A_z) randomly results in a high $\mathbf{A_z}$ value, then that feature is selected for use in the classifier despite the fact that the sample of data for this feature poorly represents its underlying distribution. Thus it is expected that if the same data employed to select the features in this manner is also employed to determine the parameters of the classifier, then bias will be introduced into the classification process because the training data poorly represents the true density of the data.

In order to analyze this bias, we simulated N features using Gaussian distributions, n of which had theoretical A_z values of $A_z^{(1)}=0.68$, and $N-n$ of which had a theoretical A_z

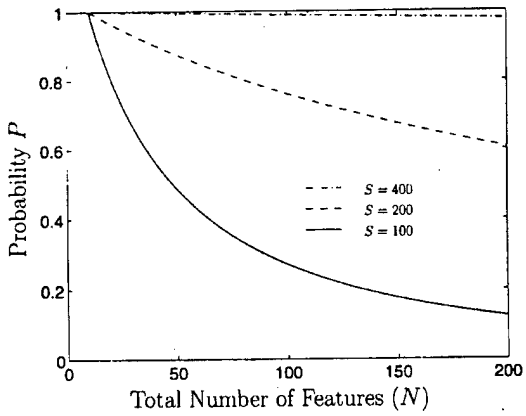


FIG. 4. A plot of the probability of selecting an optimal subset of $n=4$ features from a total of N features. There are a total of $r=10$ features with a theoretical A_z value of $A_z^{(1)}=0.7$ and $N-r$ features with a theoretical A_z value of $A_z^{(2)}=0.6$. The probability of selecting an optimal subset of features is also plotted for various sample sizes S.

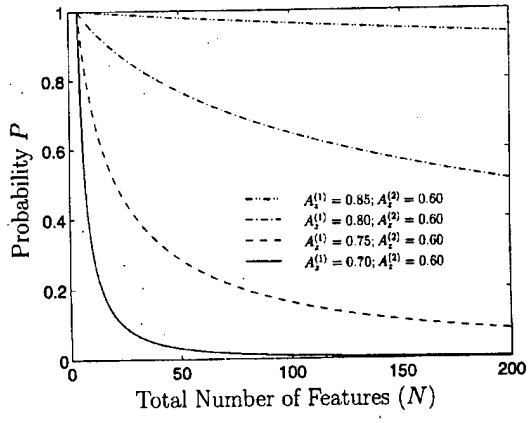


FIG. 5. A plot of the probability of selecting an optimal subset of $n=4$ features from a total of N features. There are a total of $r=4$ features with a theoretical A_z value of $A_z^{(1)}$, $N-r$ features with a theoretical A_z value of $A_z^{(2)}$ and the sample size is fixed at $S=100$. The probability of selecting an optimal subset of features is plotted at various combinations of $A_z^{(1)}$ and $A_z^{(2)}$.

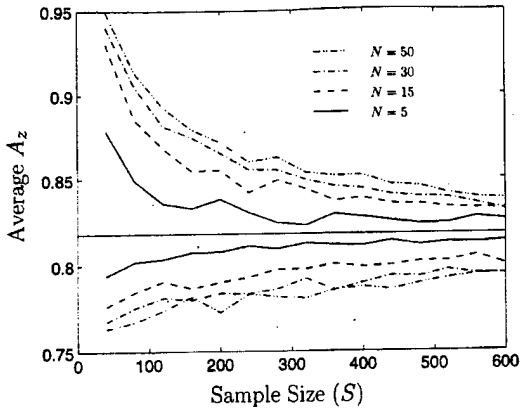


FIG. 6. A total of $r=4$ Gaussian features were simulated with a theoretical A_z value of $A_z^{(1)}=0.68$ and $N-r$ Gaussian features were simulated with a theoretical A_z value of $A_z^{(2)}=0.6$. Features were sampled 100 different times and the top $n=4$ features were selected based on the measured $\mathbf{A_z}$ values of the individual features. The selected features were then merged using linear discriminants and the training dataset and testing dataset $\mathbf{A_z}$ values were computed. The thin solid line at an A_z of 0.818 is the theoretical true A_z if the 4 independent Gaussian ($A_z=0.68$) features are merged using linear discriminants. The curves above this theoretical line are the average training dataset $\mathbf{A_z}$ values and the curves below the theoretical line are the average testing dataset $\mathbf{A_z}$ values. The same data used to select the features was employed in determining the parameters of the linear discriminants. A substantial amount of bias is introduced is one has small sample sizes and a large total number of features from which to select.

value of $A_z^{(2)}=0.60$. The n features with the highest measured $\mathbf{A_z}$ values were then combined using linear discriminant analysis to merge the n-dimensional features to a scalar decision variable representing the distance from the separation plane to the point in question. The $\mathbf{A_z}$ value of the classifier was computed using, as input to the ROC analysis, the scalar decision variable data as output from the linear discriminant classifier. The same data employed to select the n features is used to determine the parameters of the linear discriminant that merges the n features. We also tested the classifier on an independent dataset of 1000 samples. This process was repeated 100 times for each combination of parameters to obtain an average training dataset $\mathbf{A_z}$ and testing dataset $\mathbf{A_z}$ value for the classifier. Figure 6 shows a plot, for various total numbers of features N, of the average training and testing dataset $\mathbf{A_z}$ values as a function of the sample size S. The thin solid line in Fig. 6 is the theoretical true A_z value of 4 independent Gaussian features with equal variances and individual A_z values of 0.68, merged using linear discriminants. The curves above the theoretical line are the average training dataset $\mathbf{A_z}$ values, and the curves below the theoretical line are the average testing dataset $\mathbf{A_z}$ values. As this figure shows, bias is introduced when the same data are used to select and merge features. The bias is enhanced in situations where there is little data and a large total number of features N; this is the same condition in which it is the most difficult to select an optimal subset of features (see Figs. 3–5). Hence, a suboptimal subset of features is most likely selected and bias is introduced because we are employing the same dataset to select features and determine the parameters of the classifier.

To further demonstrate this, Fig. 7 plots a normalized histogram of the features that were actually selected over many trials when there were a total of 30 features ($N=30$) from which to select using sample sizes of 80 [Fig. 7(a)] and 200 [Fig. 7(b)]. The first four actually best features are more likely to be selected than any one of the remaining features, but the probability of selecting all four best features is nevertheless small. With a larger sample size the difference in the probability of selecting one of the best features versus one of the worse features is enhanced.

V. DISCUSSION

In this paper we have only dealt with a feature selection methodology that employs the performances of the indi-

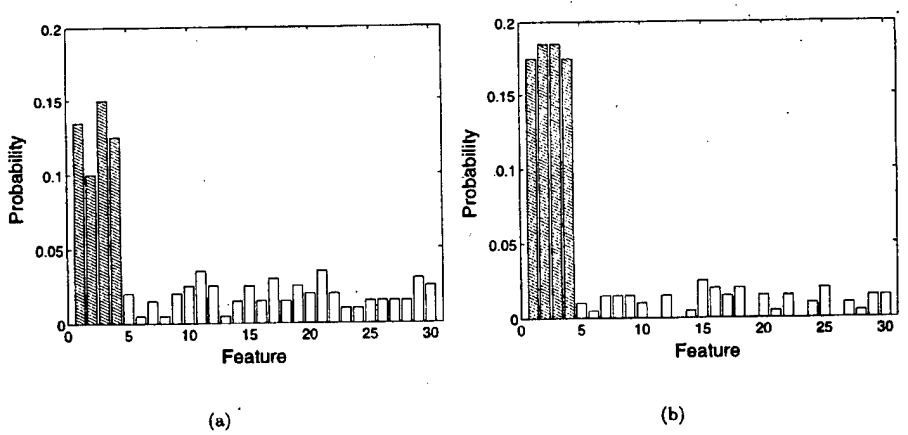


FIG. 7. Normalized histograms of the features actually selected when $r=4$, $n=4$, $A_z^{(1)}=0.68$, $A_z^{(2)}=0.6$, $N=30$ and (a) $S=80$ and (b) $S=200$. The shaded bars are the $n=4$ actually better features with a theoretical A_z value of 0.68. The better features are more likely to be selected than any one worse feature but the combined probability that one of the worse features will be selected is high. The chances are better that one will select the optimal first four features when the sample size is large (b).

vidual features to select the subset to be merged using a classifier. In this situation there are a total of N measured $\mathbf{A_z}$ values to analyze. Many other feature selection algorithms such as stepwise feature selection[9] and genetic algorithm feature selection[7,10,11] measure the performance of combinations of features instead of individual features. In these types of algorithms instead of having a total of N measurements, there is a total of $N!/n!(N-n)!$ possible measurements. This situation is possibly even more problematic and more sensitive to the distribution of measured $\mathbf{A_z}$ values than the individual feature selection method we have studied, because the probability that a theoretically poor subset of features will yield a high performance measure is increased.

We have used and merged only independent features. More information is gained when two independent features are merged instead of two correlated features. It should be noted that the optimal classifier for n independent Gaussian distributed features with equal variances is a linear discriminant, which provides justification for the method we have chosen to merge our simulated features. In general, feature selection methods that measure the performances of combinations of features must take into account correlations among the features. If the features in a similar study were correlated, then a feature selection methodology which employs the performances of the individual features would clearly be suboptimal. We have restricted our attention to the simpler case of independent features here because it provides a basis for analyzing the difficulty in selecting features and merging the selected features.

We have chosen the area under the maximum likelihood estimate of the ROC curve, A_z,[14] as our performance measure in selecting the subset of features. Because we have limited data, there is a distribution associated with the measured $\mathbf{A_z}$ values, and we have studied the difficulty in selecting and merging features using $\mathbf{A_z}$ as the selection criterion; in practice, there are always problems in representing the performance of classifiers by a single scalar measure.[20] In these simulation studies our features had equal variances. Under this condition there is theoretically no ambiguity in comparing A_z values only, because the population ROC curves will never cross if they have unequal A_z values.[12,13]

In practice, however, one could run into difficulty when using A_z as a performance measure because two very different ROC curves could have equal A_z values. Any scalar performance measure, however, is going to have a distribution associated with it due to limited data and is going to have difficulty with respect to the ambiguity associated with representing the performance of a classifier by a scalar quantity and, hence, the results presented in this paper can be extended, in general, to different classifier performance measures. Equation (3) is a general equation, and similar equations could be derived for different performance measures but with different limits of integration and density functions. The datasets employed in this study had equal numbers of abnormal and normal cases, i.e., $S_a=S_n$. One can study the effect of unequal abnormal and normal cases by noting that only the variance in the measured $\mathbf{A_z}$ value changes [see Eq. (2)].

VI. CONCLUSIONS

We have shown that when one has small sample sizes and a large total number of features from which to select, the probability of selecting an optimal or even near-optimal subset of features is small. As the sample size increases and the total number of features decreases, the probability of selecting an optimal subset approaches 1. The difficulties of feature selection, however, are twofold because bias is also introduced if one selects features and determines the parameters of a classifier using the same dataset. This bias is caused by the fact that one is preferentially selecting features that misrepresent their underlying density functions. The bias is greater when one has small sample sizes and a large total number of features, the same situation in which it is unlikely that one will select an optimal subset of features.

The results in this work have dealt with a fairly simple and ideal situation of independent features that, in the limit as the number of samples goes to infinity, are optimally merged using linear discriminants. In practice, the difficulties of features selection are, in fact, more troublesome than presented here. The results presented can be extended to performance measures other than A_z and inferences can be made about other feature selection methods that combine features instead of measuring their individual performances.

APPENDIX

Assume N independent observations x_i with a joint density function $f_1(x_1)f_2(x_2)\cdots f_N(x_N)$ and joint distribution function $F_1(x_1)F_2(x_2)\cdots F_N(x_N)$. Define the set of all observations as $\mathcal{S}=\{1,2,...,N\}$ and let $\mathcal{S}$ be partitioned into two subsets $\mathcal{B}$ and $\mathcal{W}$. We will now focus attention on one element of the partition $\mathcal{B}$, which will be labeled x_b. The probability (i) that x_b falls between x and $x+dx$ is $f_b(x)dx$, (ii) that all remaining observation in $\mathcal{B}$ are larger than x is $\Pi_{i\in\mathcal{B};i\neq b}(1-F_i(x))$, and (iii) that all observations in $\mathcal{W}$ are less than x is $\Pi_{i\in\mathcal{W}}F_i(x)$. The probability that (i), (ii), and (iii) all occur simultaneously is

$$f_b(x)dx\prod_{i\in\mathcal{W}}F_i(x)\prod_{i\in\mathcal{B};i\neq b}(1-F_i(x)). \tag{A1}$$

Integrating the above expression over all x arrives at the total probability that the observations in $\mathcal{W}$ are less than observation x_b which is less than all other observations in $\mathcal{B}$. By summing over all of the elements in $\mathcal{B}$, one arrives at the probability that all of the observations in $\mathcal{B}$ are larger than all of the observations in $\mathcal{W}$, i.e.,

$$P(x_{\{\mathcal{B}\}}>x_{\{\mathcal{W}\}})=\sum_{j\in\mathcal{B}}\left[\int dx f_j(x)\times\prod_{i\in\mathcal{W}}F_i(x)\prod_{i\in\mathcal{B};i\neq j}(1-F_i(x))\right]. \tag{A2}$$

If the n observations of $\mathcal{B}$ are sampled from the same density function $f_1(x)$, $r-n$ of the observations in $\mathcal{W}$ are also sampled from $f_1(x)$, and the remaining $N-r$ observations have common density $f_2(x)$, then Eq. (A2) simplifies to

$$P(x_{\{\mathcal{B}\}}>x_{\{\mathcal{W}\}})=n\int dx f_1(x) \times F_1(x)^{r-n}F_2(x)^{N-r}(1-F_1(x))^{n-1}. \quad \text{(A3)}$$

The above expression assumes that the observations in $\mathcal{B}$ are a specific subset of the r observations with density $f_1(x)$. There are, however, a total of $r!/n!(r-n)!$ possible subsets that contain n observations with density $f_1(x)$. Multiplying Eq. (A3) by this fraction, we arrive at the probability that the n largest observations all come from the density function $f_1(x)$,

$$\frac{r!}{(n-1)!(r-n)!}\int dx f_1(x)F_1(x)^{r-n} \times F_2(x)^{N-r}(1-F_1(x))^{n-1}. \quad \text{(A4)}$$

If our observations $\mathbf{x}$ are measured $\mathbf{A_z}$ values, then the above expression becomes Eq. (3).

ACKNOWLEDGMENTS

The authors thank Mark A. Anastasio for his helpful discussions and Darrin C. Edwards for his help with order statistics. This work was supported in parts by grants from the US Army Medical Research and Materiel Command (DAMD 17-96-1-6058 and 17-97-1-7202) and USPHS Grant No. RR11459. Maryellen L. Giger is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

[1] D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swetts, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).
[2] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).
[3] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).
[4] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," Acad. Radiol. **6**, 22–33 (1999).
[5] Y. Wu, K. Doi, C. E. Metz, N. Asada, and M. L. Giger, "Simulation studies of data classification by artificial neural networks: Potential applications in medical imaging and decision making," J. Digital Imaging **6**, 117–125 (1993).
[6] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancers," Radiology **187**, 81–87 (1993).
[7] B. Sahiner, H.-P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Gootsitt, "Image feature selection by a genetic algorithm: Application to classification of mass an normal breast tissue," Med. Phys. **23**, 1671 (1996).
[8] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," IEEE Trans. Pattern. Anal. Mach. Intell. **19**, 153–158 (1997).
[9] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice-Hall, New York, 1992).
[10] M. A. Kupinski, M. L. Giger, and K. Doi, in *Digital Mammography*, edited by K. Doi, International Congress Series. pp. 401–404 (Elsevier, New York, 1996).
[11] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," Pattern Recogn. Lett. **10**, 335–347 (1989).
[12] C. E. Metz, "Basic principles of ROC analysis," Semin. Nucl. Medi. **VIII**, 283–298 (1978).
[13] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).
[14] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest. Radiol. **24**, 234–245 (1989).
[15] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745–750 (1996).
[16] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (Springer-Verlag, New York, 1996).
[17] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29–36 (1982).
[18] H. A. David, *Order Statistics* (Wiley, New York, 1970).
[19] J. D. Gibbons, I. Olkin, and M. Sobel, *Selecting and Ordering Populations: A New Statistical Methodology* (Wiley, New York, 1977).
[20] M. A. Anastasio, M. A. Kupinski, and R. M. Nishikawa, "Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach," IEEE Trans. Med. Imaging **17**, 1089–1093 (1998).